Universal Dependencies for Mandarin Chinese

Rafaël Poiret[a] · Wong Tak-Sum[b] · John Lee[c] · Kim Gerdes[d] · Herman Leung[c]
[a] Tsukuba University, Faculty of Humanities and Social Sciences, Tsukuba (Japan),
[b] The Hong Kong Polytechnic University, Department of Chinese and Bilingual Studies (Hong Kong S.A.R.),
[c] City University of Hong Kong, Department of Linguistics and Translation (Hong Kong S. A. R.),
[d] Paris-Saclay University, Lisn (CNRS), Paris (France),
Corresponding author email address: poiret.rafael.gf@u.tsukuba.ac.jp

Abstract This article presents a Universal Dependency (UD) annotation scheme for Mandarin Chinese, as well as the current UD Chinese HK treebank. Our focus is mainly on parts-of-speech (POS) tags and syntactic relations, with a quite large array of phenomena investigated. The main goal is to make transparent the linguistic consideration behind our annotation choices, and show how we articulated these choices with the criteria of Universal Dependencies. This scheme has been developed with reference to two other dependency schemes for this language, i.e. the Chinese Stanford Dependencies (Chang et al. 2009) and the Chinese Dependency Treebank (HIT-SCIR 2010). We provide mappings between our scheme and the two others. The content of the UD Chinese HK treebank is discussed in relation to the other UD treebanks for Chinese, and the inter-annotator agreement on POS and dependency annotation is reported. Our proposed scheme is motivated by reasoned linguistic analysis, is suitable for cross-linguistic comparison, and produced a high level of agreement between annotators.

Keywords Mandarin · Universal Dependencies · Treebank · Annotation scheme

Funding
This work was supported by a grant from the PROCORE-France/Hong Kong Joint Research Scheme sponsored by the Research Grants Council and the Consulate General of France in Hong Kong (Reference No.: F-CityU107/15 and No. 35322RG); and by a grant (Project No. 9360163) from the Hong Kong Institute for the Data Science at the City University of Hong Kong.

Conflicts of interest
The authors declare they have no conflict of interest.

Availability of data and material
The UD Chinese HK treebank data is available at:
https://github.com/UniversalDependencies/UD_Chinese-HK/

1 Introduction

The Universal Dependencies (UD) project (de Marneffe et al. 2014; Nivre et al. 2020) constitutes an important homogenization effort to synthesize ideas and experiences from different dependency treebanks in different languages. The aim of the UD project is to facilitate multilingual research on syntax and automatic parsing by proposing a unified annotation scheme for all languages. The scheme has triggered some debate on the syntactic foundation of some choices that have been made (Osborne and Gerdes 2019). UD contributors are invited to find some compromises between the six following criteria[1]:

1. UD needs to be satisfactory on linguistic analysis grounds for individual languages;

[1] https://universaldependencies.org/introduction.html

2. UD needs to be good for linguistic typology, i.e. providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families;
3. UD must be suitable for rapid, consistent annotation by a human annotator;
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a habitable design, and it leads us to favor traditional grammar notions and terminology;
5. UD must be suitable for computer parsing with high accuracy;
6. UD must support well downstream understanding tasks (relation extraction, reading comprehension, machine translation, …)

However, some of these criteria are necessarily contradictory. Indeed, it is felt that syntactic correctness, applicability of the schemes for NLP tools and purposes, and above all universality, cannot all be fulfilled at the same time (Gerdes and Kahane 2016). Although no separate explicit annotation scheme exists for most UD treebanks, universality seems to outweigh other considerations, as exemplified by the importance of semantic criteria such as the central distinction between lexical words and function words, a choice that results in more similar structures among languages than a purely distributional approach would provide. Meanwhile, a revision of UD towards principles of contemporary linguistic typological theory has been proposed (Croft et al. 2017), and UD has recently been defended as a coherent (monostratal) theory for the annotation of typologically diverse languages in (de Marneffe et al. 2021).

Despite the drawbacks, the project encountered a tremendous success. As of the last version released (version 2.8), more than four hundred contributors around the world have produced 202 treebanks for 114 languages. A UD scheme for Mandarin Chinese (hereafter, Chinese) has been developed at the City University of Hong Kong since 2016. In this context, Leung et al. (2016) highlighted several advantages and disadvantages of the choices made in their adoption of UD for Chinese. The gaps and problems described show more generally that morphosyntactic categories that were originally created for Indo-European languages require considerable adaptation efforts so that the linguistic features of Mandarin Chinese could be integrated into UD. Some of these problems can be solved by a greater universality of the vocabulary used to describe the syntactic distinctions.

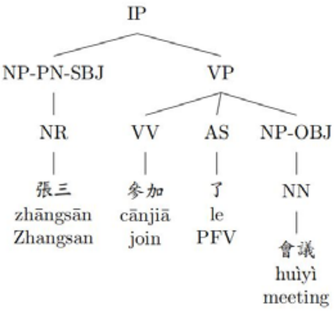
This paper presents an in-depth and more complete description of the annotation choices we made for the creation of our UD treebank for Chinese. Both the guidelines and the treebank are called UD Chinese HK[2]. A direct benefit of this work is to ease the harmonization between our choices and those made in the other UD treebanks for Chinese, but also for other languages. Our focus is mainly on the language-specific linguistic considerations that underlie the creation of the treebank, rather than on the computing aspects. We show in each subsection how we met the different criteria posited by UD, sometimes requiring trade-offs. The article starts out with a brief overview of existing dependency annotation schemes for Chinese and how they compare overall to the UD scheme. Section 3 describes word segmentation differences between Penn Chinese and Chinese UD. We describe a few of the Chinese POS tag choices of our scheme in section 4. Section 5 is devoted to how UD relations are handled for Chinese. Section 6 discusses Chinese compounds and constructions. Section 7 discusses the content of the UD Chinese HK treebank in relation to the other UD treebanks for Chinese, and reports the inter-annotator agreement on POS and dependency annotation. Section 8 concludes and proposes future work. We make explicit how each UD criterion was taken into account. The Chinese characters that we used in this paper are in traditional Chinese (漢字 *hànzì*). Traditional Chinese characters are used in Taiwan, Hong Kong and Macau. In contrast, simplified Chinese characters (汉字 *hànzì*), which have been introduced more recently by the government of the People’s Republic of China, are used in Mainland China, Malaysia, and Singapore.

[2] At the time this paper is written, the guidelines (the documentation of tags, features, and relations) uploaded to the official website of Universal Dependencies for Chinese correspond to the UD Chinese HK presented in this paper.

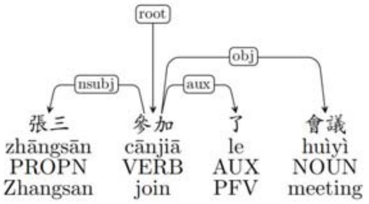
2 Syntactic annotation schemes for Mandarin Chinese

There are globally two types of syntactic treebanks, according to whether they follow constituency grammar or dependency grammar approaches. One key characteristic of constituent treebank annotation is that smaller syntactic units (tokens) are grouped together into larger constituents (phrases). Those constituents further group together to form even larger syntactic constituents (larger phrases or clauses), as exemplified in (1a). In contrast, in the dependency grammar approach, the only nodes in the tree are the tokens themselves. The relations between tokens are expressed in terms of governor-dependent links graphically represented as arrows, as shown in (1b). Although each governor can have multiple dependents, in most basic implementations of dependency treebanks, a dependent has only one governor except for the root node, giving a tree structure.

(1a) Constituent tree (CTB)



(1b) Dependency tree (UD)



'Zhangsan joined a meeting.'

The Chinese Treebank (CTB) is one of the most well-known Chinese constituent treebanks. Its guidelines are modeled after, and heavily influenced by, the Government and Binding theory introduced in (Chomsky 1981), although the framework is not adopted wholesale (Xue et al. 2000). CTB started at the University of Pennsylvania, before moving to the University of Colorado at Boulder, and is now maintained at Brandeis University. The current version of the treebank contains over 2 million words comprised of data from newswire, broadcast material, web text, and transcribed speech (Xue et al. 2013). The treebank continues to use the guidelines for word segmentation, tagging, and bracketing that were last developed in 2000 during its development at the University of Pennsylvania. Thus, in the rest of this paper, we will refer to the corpus itself as CTB, but to its annotation scheme as Penn Chinese. Two widely used Chinese dependency schemes are Stanford Dependencies for Chinese (hereafter, Stanford Chinese), developed by Huihsin Tseng and Pi-Chuan Chang (see Chang 2009; Chang et al. 2009), and the Chinese Dependency Treebank (CDT) developed by the Harbin Institute of Technology Research Center for Social Computing and Information Retrieval (see Che et al. 2012; HIT-SCIR 2010).

Stanford Chinese adopts its part-of-speech (POS) tagset directly from Penn Chinese, instead of establishing its own. The differences between constituent and dependency principles do not directly affect tagging nor word segmentation. Stanford Chinese and Penn Chinese relations and structures are closer to UD, and above that Stanford-Chinese-to-UD transformation tools already exist[3]. Therefore, we have taken many elements primarily from Stanford Chinese and CTB. However, we have simultaneously made some choices that differ from some traditional Chinese linguistics analyses which Stanford Chinese and CDT follow. At the macroscopic level, our implementation of Chinese UD differs from the other two dependency schemes in how the number of POS tags and the number of relations are distributed, as summarized in Table 1.

[3] https://nlp.stanford.edu/software/stanford-dependencies.html

Table 1. Summary comparison of dependency schemes

	Stanford Chinese	**CDT (Harbin)**	**UD**	**UD Chinese HK**
POS tags	33	26	17	Identical to UD
Dependency relations	45	15	37	36 of the 37 standard UD relations; added 16 language-specific relations

Albeit due to UD restrictions, the much smaller set of POS tags in UD Chinese HK is compensated by a greater number of dependency relations. The idea is that more specific (sub)categories of POS should be recoverable with the relation labels. UD also provides the option of specifying additional morphosyntactic information on words, called "features". The information stored in this layer of annotation ranges from case to noun class, tense, modality, number, and so on. Any information that might be lost from UD's smaller, fixed tagset can thus be relegated to feature annotation, if they are not also recoverable from the relation annotation. While Stanford Chinese has the highest amount of overlapping POS and dependency relations among the three schemes, CDT takes the opposite approach of simplifying both the POS tags and dependency relations. Collapsing CDT's 8 noun POS categories into just two (nouns and proper nouns), CDT would have only three more POS tags than UD Chinese HK. In the following section, we first address how we deal with word segmentation. The rest of the paper is devoted to the more salient issues we encountered in developing UD Chinese HK with regard to POS tagging and syntactic annotation.

3 Word segmentation

Word segmentation is not a trivial issue in Chinese since word boundaries are not delimited in the writing system. We adopt — with some minor differences to be explained — the Penn Chinese guidelines for segmentation (Xia 2000a)[4]. These guidelines have already mapped out a comprehensive treatment of both general and specific cases.

The Penn Chinese guidelines notion of "word" based on a minimal syntactic unit (Xia 2000a) is followed. UD similarly proposes that the basic units of annotation should be syntactic words rather than phonological or orthographic words (Nivre et al. 2016). In effect, Penn Chinese regards many bound morphemes (such as aspect markers) as individual tokens, while treating less productive ones (such as the plural 們 *mén* for pronouns and human nouns) as part of a token (although productivity is not in itself a reliable test of wordhood). Treating all bound morphemes as part of another token may be a more consistent treatment. However, we also find this approach impractical for a language whose morphemes are already represented by characters that are immutable in their written form, regardless of their phonological or morphosyntactic realization. This is in contrast to many Indo-European and other languages where, for example, conjugated verb forms can be so different from their stem or root form that they look like completely different words (such as *ir* 'to go' and *fui* 'I went' in Portuguese). In Chinese, neither the aspect markers nor the verbs they follow ever change form, no matter what other morphemes or words are next to them (我 *wǒ* 'I' + 看 *kàn* 'see' + 過 *guò* 'experiential aspect marker' = 'I have seen').

Our word segmentation guidelines differ from those of Penn Chinese only in that we systematically split into separate tokens all verb-verb compounds (e.g., 摔 *shuāi* 'fall' and 破 *pò* 'break' in 我 摔 破 了 腿 *wǒ shuāi pò le tuǐ* 'I fell breaking my leg.') and verb-x-verb compounds (where x stands for the affirmative potential marker 得 *dé* or negative potential marker 不 *bù* in 我 找 不 到 *wǒ zhǎo bú dào* 'I can't find'). Penn Chinese, on the other hand, separates or combines them depending on at least two different criteria: the

[4] We were not able to obtain segmentation guidelines for the CDT to make a comparison.

number of syllables and semantic (de)compositionality. One reason we diverge from Penn Chinese is to minimize the processing time needed for employing human annotators to decide whether a given compound fits one criterion or the other, which could cause disagreement between annotators. It also reduces the need for employing a human annotator to change what can be automatically tokenized apart, particularly in the case of the verb-x-verb compounds. Lastly, given how productive the verb-verb compound is, and its separability, it makes more sense to tokenize it apart rather than to treat each possible combination as a non-compositional unit.

4 Parts of speech

UD Chinese HK uses all of UD's 17 parts-of-speech (UDPOS) tags (Nivre et al. 2016). We adopt heavily from the Penn Chinese POS tagset (Xia 2000b). Our tags for adverb (`ADV`), coordinating conjunction (`CCONJ`), interjection (`INTJ`), pronoun (`PRON`), proper noun (`PROPN`), punctuation (`PUNCT`) and subordinating conjunction (`SCONJ`) correspond exactly to their counterpart in the Penn tagset. The correspondence between the POS tagset of our UD Chinese HK and that of Penn Chinese and CDT is also presented in Appendix 1. However, our scheme still differs from the Penn POS system in a few places, since UD's tagset is smaller and does not correspond neatly to all of Penn Chinese's tags. Since UD does not allow sub-typing of POS tags or the creation of language-specific tags, we adhere to this restriction. The list of UDPOS along with some Chinese examples are presented in Table 2.

Table 2. Parts-of-speech tags in Chinese UD v2 with examples

UDPOS	Description	Examples in UD Chinese HK
`ADJ`	Adjective	Attributive adjective: 香 花 *xiāng huā* 'fragrant flower' Predicative adjective: 花 很 香 *huā hěn xiāng* 'the flower is very fragrant' Ordinal numbers: 第二 *dìèr* 'second'
`ADP`	Adposition	Prepositions: 在 *zài* 'at'; 對 *duì* 'to(ward)' Postpositions: 之上 *zhīshàng* 'above'; 內 *nèi* 'inside' Valence markers: 被 *bèi*; 把 *bǎ*
`ADV`	Adverb	很 *hěn* 'very'; 也 *yě* 'also'; 經常 *jīngcháng* 'often'
`AUX`	Auxiliary	Passive auxiliary: 被 *bèi* Modal auxiliary verbs: 可以 *kěyǐ* 'can'; 會 *huì* 'will' Aspect markers: (沒)有 (*méi*)*yǒu* (negative perfective); 着 *zhe* (durative); 了 *le* (perfective); 過 *guò* (experiential)
`CCONJ`	Conjunction	和 *hé* 'and'; 或 *huò* 'or'; 但 *dàn* 'but'
`DET`	Determiner	這 *zhè* 'this'; 那 *nà* 'that'; 前 *qián* 'the previous'
`INTJ`	Interjection	哦 *ó* 'oh'; 哎喲 *āiyō* 'aiyo'
`NOUN`	Noun	Nouns: 草 *cǎo* 'grass'; 今天 *jīntiān* 'today' Classifiers: 個 *gè* (generic classifier)
`NUM`	Numeral	五 *wǔ* 'five'
`PART`	Particle	Genitive: 的 *de* Sentence-final particles: 嗎 *ma* (question particle)
`PRON`	Pronoun	我 *wǒ* 'I'; 他 *tā* 'he'; 這 *zhè* 'this'
`PROPN`	Proper noun	歐陽修 *Oūyáng Xiū* 'Ouyang Xiu', 艾恩斯坦 'Einstein'
`PUNCT`	Punctuation	。(period); 《》(title quotation marks)
`SCONJ`	Subordinating conjunction	如果 *rúguǒ* 'if'; 雖然 *suīrán* 'although'; 的話 *dehuà* 'if'
`SYM`	Symbol	© (copyright symbol); * (asterisk); ☺(emoji)

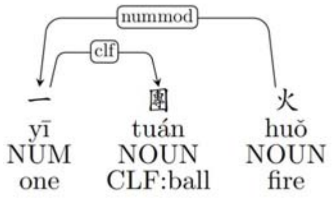
VERB	Verb	Copula: 是 *shì* 'be' (in its non-copular meanings/functions) Possessive: 有 *yǒu* 'have, exist' Other: 吃 *chī* 'eat'
X	Other	Foreign word with unknown part of speech

Note that not all POS tags correspond neatly and completely between the different schemes. Some tags cover other categories not included in the corresponding UD Chinese HK tag (indicated by double parentheses in Appendix 1). For example, the UD Chinese HK tag AUX includes the passive auxiliary 被 *bèi*, modal auxiliary verbs, and aspect markers. We treat as modal and modal-like auxiliary verbs the verbs that can be pre-modified by the negator 不 *bù* but cannot take an object and be post-modified by an aspect marker. In both Penn Chinese and CDT, modal auxiliary verbs are in the same category as regular verbs—VV 'other verbs' and v 'verb', respectively. In other words, in UD Chinese HK, the tag VERB covers a different set of verbs than Penn Chinese VV or CDT v. Another example is that CDT does not differentiate between coordinating and subordinating conjunctions, grouping them both under c 'conjunction', while UD Chinese HK separates them between CCONJ and SCONJ, and Penn Chinese between CC and CS, respectively. This means that conversion from one system to another will have to include exceptions. There are three linguistic categories — namely, predicate adjectives (tagged as ADJ) and postpositions (tagged as ADP), and classifiers (tagged as NOUN) — for which our treatment differs significantly from one or both of the previous treebanks mentioned here. Our analysis of adjectives has already been presented in (Leung et al. 2016). Since the publication of that paper, we made some changes concerning the analysis of classifiers. Additionally, objective criteria to define adpositions in Chinese have been found necessary for the annotation of POS tags. The second and third issues are therefore discussed in the subsections below.

4.1 Classifiers

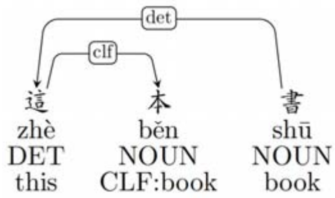
Classifiers (more specifically, "numeral classifiers" per Aikhenvald 2000) are an indispensable lexical category in Chinese as well as in many East Asian and Southeast Asian languages. In Chinese, they are often obligatory in a noun phrase with a numeral modifier (2) and optional with a demonstrative (3). The classifier is attached to the numeral if a numeral is present, or else to the determiner. This is because typologically, classifiers group with the modifier rather than the noun; there are Noun Num Clf languages, but not Clf Noun Num languages (Croft, personal communication). In both cases, the relevant dependency relation is clf.

(2)



'a ball of fire'

(3)

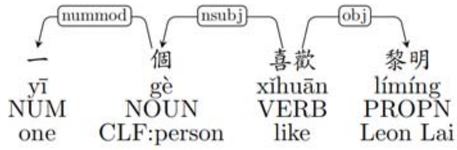


'this book'

Classifiers can also referentially substitute the noun, and then they function as the head of the phrase, with the modifiers congruously attached to them (4). They may also appear in an indefinite noun phrase in object position without a numeral or demonstrative (5). Then, they are considered to function as an indefinite determiner, and annotated with det[5].

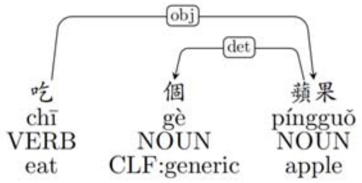
[5] Our analysis of classifiers is based on the proposition made by William Croft, in a discussion conducted on the Github of Universal Dependencies.

(4)



‘one likes Leon Lai’

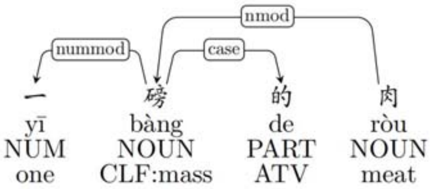
(5)



‘eat an apple’

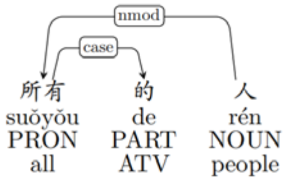
Lastly, when the genitive 的 *de* is inserted in between the classifier and the noun, the classifier is treated as the governor of the numeral or ordinal number, and it is labeled as an nmod dependent of the noun (6). Note that the same policy is followed for pronoun that can function as determiner (e.g. 一切 *yīqiè* ‘all’; 所有 *suǒyǒu* ‘all’; 全部 *quánbù* ‘all’) when they are separated from the noun by 的 *de* (7).

(6)



‘a pound of meat’

(7)



‘all people’

It is likely due to the unique syntactic distribution of classifiers that both Penn Chinese and CDT give them unique POS tags — M for measure word and q for quantity, respectively. Since UD does not allow language-specific POS tags to be created for language-specific categories, classifiers must be merged with an existing tag. As of UD v2, they are tagged NOUN. The choice to tag them as NOUN fits well with the syntactic distribution of Chinese classifiers in that they can function as heads of noun phrases (as shown in (5)). Besides, they are similar to measure words in languages such as English (e.g., *a head of cattle*), Danish (e.g., *en kop kaffe* ‘a cup of coffee’) and French (e.g. *un kilo de viande* ‘a kilo of meat’) where such measure words are nominal in nature. Still, classifiers could be tagged PART when their role is more functional than referential. This occurs in every other case mentioned above except when it serves as the head of a noun phrase. Therefore, UD’s choice of grouping classifiers with nouns in the POS tag categorization is a compromise for the sake of comparability with non-classifier languages. That is, we meet the UD criterion 2: “UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families”. Nonetheless, their classifier status is preserved in the feature column as NounType=Clf.

4.2 Adpositions

In Chinese, the delimitation is not trivial between prepositions and verbs, and neither between postpositions and location nouns. As for the former, this is explained by the numerous pairs of homophones (e.g. verb 在 *zài* ‘be.at’ vs. preposition 在 *zài* ‘at’, verb 給 *gěi* ‘give’ vs. preposition 給 *gěi* ‘to’). For a consistent annotation of the corpus, objective criteria to define both prepositional and postpositional morphosyntactic categories are thus crucial. This is what we tackle in the present demonstration, based on (Paul 2015). By doing so, we meet criterion 1: “UD needs to be satisfactory on linguistic analysis grounds for individual languages”. Contrarily to verbs, prepositions in Chinese cannot be modified by adverbs or negated. This fact becomes visible once the prepositional phrase candidate is fronted (8a), as in example (8b) which contains the adverb 已經 *yǐjīng* ‘already’.

(8)	a.	給	瑪麗，	我	已經	發	了。
		gěi	*mǎlì*	*wǒ*	*yǐjīng*	*fā*	*le*
		to	Mary	I	already	send	SFP

‘To Mary, I have already sent it.’

b.	*已經	給	瑪麗，	我	發	了。
	yǐjīng	*gěi*	*mǎlì*	*wǒ*	*fā*	*le*
	already	to	Mary	I	send	SFP

‘To Mary, I have already sent it.’

Unlike verbs, prepositions cannot function as predicates (9), and require their complement to be overt (10).

(9)	*瑪麗	從	香港。
	mǎlì	*cóng*	*xiānggǎng*
	Mary	from	Hong Kong

‘Mary is from Hong Kong.’

(10)	瑪麗	在	*(家)	睡覺。
	mǎlì	*zài*	*jiā*	*shuìjiào*
	Mary	at	home	sleep

‘Mary sleeps at home.’

Paul (2015: 55-57) provides a (non-exhaustive) list of prepositions, with two subcategories: prepositions having no verbal counterpart, and prepositions having a verbal counterpart. However, the list does not contain unclear cases like 趁 *chèn* ‘taking advantage’ vs. 趁 *chèn* ‘while’. This list should be completed by a re-examination of Chao’s (1968: 754-769)’s list with the criteria of Paul (2015) aforementioned. Paul (2015: 95-97) also provided a list of postpositions. She proposed two criteria for postpositions. Postpositions (11a), unlike location nouns (11b), do not accept the subordinator 的 *de*. Postpositions, like prepositions, require their complement to be overt (12).

(11)	a.	桌子	(*的)	上
		zhuōzi	*de*	*shàng*
		table	ATV	on

‘on the table’

(12)	*(新年)	以前	走
	xīnnián	*yǐqián*	*zǒu*
	new.year	before	leave

‘Leaving before the new year.’

b.	桌子	的	上邊
	zhuōzi	*de*	*shàngbian*
	table	ATV	upper.side

lit. ‘the top of the table’

Postpositions are annotated with the relation `case:loc`. On top of these prepositions and postpositions, we also treat valence markers 把 *bǎ* and 被 *bèi* as adpositions. This analysis is motivated to maintain consistency with Stanford Chinese and CDT. In Stanford Chinese, both markers are treated in syntax like prepositions, although two specific relations are used. In CDT, 把 *bǎ* is treated as preposition, and 被 *bèi* as an auxiliary. Here, a balanced analysis is found where both 把 *bǎ* and 被 *bèi* are tagged as prepositions, except for 被 *bèi* which is tagged as an auxiliary in the short passive construction. More details will be given in the next section on syntactic relations.

5 Syntactic relations

Our adoption of UD for Chinese has presented a number of challenges, with regard to syntactic relations. Some are due to particular constructions whose analyses are controversial or under-researched. Others are due to what are potential gaps in the UD design thus far. We discuss these issues in the subsections below. We use 36 of the 37 syntactic relations available in UD as laid out on the official UD website (de Marneffe et al. 2016), leaving out `expl` since expletives do not exist in Chinese. For the correspondence between Chinese UD relations and those of Stanford Chinese and CDT, the readers are referred to Appendix 2. We indicate in Appendix 2 that some relations in Stanford Chinese and CDT may cover usage cases (those with double parentheses) not covered by the corresponding Chinese UD relation. For instance, Chinese UD and Stanford Chinese both have the specific relations `ccomp` for clausal complements. On the other hand, CDT labels all object arguments of a verb, whether a noun phrase or a clause, as `VOB`. Therefore this CDT relation is found corresponding to both `ccomp` and `obj` relations in Chinese UD. Appendix 2 also notes those relations in Stanford Chinese and CDT that correspond to only a part of the possible usage cases covered by the corresponding Chinese UD relation (those with the "less than" symbol <). For example, the Chinese UD label `appos` is used for appositional phrases whether they are offset by commas, parentheses, or juxtaposed immediately after a noun phrase (e.g., 'John, *the man who visited me yesterday*, was blind'). If the italicized appositional phrase were in parentheses instead, then the Stanford Chinese relation `prnmod` can be used to link it to the preceding noun *John*. Hence, Stanford Chinese `prnmod` covers only one of the possible usage cases addressed by Chinese UD `appos`. Additional arrow symbols — ↑ and ↓ and ↺ — are used to indicate that either the head is different, the dependent is different, or the head-dependent direction is reversed, respectively. The 16 language-specific dependency relations, as permitted by UD, are shown in Appendix 3. Language-specific relations are labeled as subcategories by adding a colon and an additional label.

5.1 Defining syntactic relations in Universal Dependencies

More often than not in UD-based research papers and documentation, explicit criteria to define even the most core syntactic relations are not provided. We believe this situation greatly impacts the consistency of annotation between the different languages treebanks. The necessary preliminary discussion is done in this subsection, based on (Andrews 2007) and (Zeman 2017).

The design of the UD syntactic relations is grounded on the distinction between core arguments and obliques. An argument is defined according to the notion of predicate. The predicate defines a type of situation which itself entails various semantic roles (ways of participating in that situation). Arguments fulfill these roles. On the other hand, modifiers (also called adjuncts) bear semantic roles not entailed by the situation expressed by the predicate. They typically express the circumstance of the action, such as the time or location. The two most fundamental semantic roles in all languages are the ones of agent (entity responsible for an action) and patient (entity affected by an action). In a sentence, the predicate is typically expressed by a verb. Two-argument verbs requiring these two semantic roles are called Primary Transitive Verbs. These two roles are expressed in a standard way in every language. There are basically three techniques, called coding properties, used for this purpose: linear placement, agreement and case-marking. The same grammatical treatment in one language can be applied to arguments expressing semantic roles other than agent and patient. The argument of a transitive verb that gets the same grammatical treatment as an agent of a Primary Transitive Verb has the grammatical function A. The argument of a transitive verb that gets the same grammatical treatment as a patient of a Primary Transitive Verb has the grammatical function P. Finally, the argument in intransitive sentences (where the verb has an A or P argument missing) that takes the same grammatical treatment as the single argument of a one-argument predicate has the grammatical function S. In UD, arguments that have one of the S/A/P functions are core arguments. A and S arguments are subjects. P arguments are objects. As mentioned earlier, UD distinguish between clausal and nominal dependents. Hence, nominal subjects and nominal objects are labeled `nsubj` and `obj` respectively; and clausal subjects and clausal objects are labeled `csubj` and `ccomp` respectively.

5.2 Subjects and objects

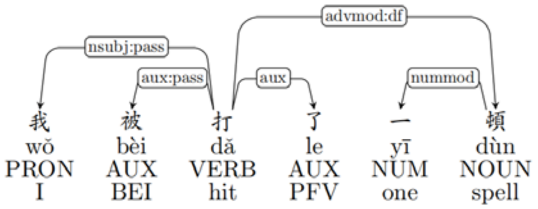
The major syntactic relations, namely, the subject and the object, depend heavily on word order in Chinese (Lu et al. 2015). As an SVO language, in a prototypical transitive sentence, subject and object often occupy the pre-verbal and post-verbal positions respectively (Peck and Lin 2019). Other word orders, like OSV, are communicatively marked (Tremblay 2005; Tremblay and Beck 2013). Example (13) shows a typical SVO example where 吃 *chī* 'eat' is the verb, while 張三 *zhāngsān* is the agent subject and 一隻蘋果 *yī zhī píngguǒ* 'an apple' is the patient object.

(13)	張三	吃	一	隻	蘋果。
	zhāngsān	*chī*	*yī*	*zhī*	*píngguǒ*
	Zhangsan	eat	one	CLF	apple

'Zhangsan eats an apple.'

The patient subjects of passive constructions are annotated with `nsubj:pass`. When the agent is not expressed, the morpheme 被 *bèi* that we normally treat as a preposition (see subsection 5.3.2 below), is analyzed with the relation `aux:pass`, as illustrated in (14). A motivation underlying our choice to treat 被 *bèi* as an auxiliary in those cases is that the relations `aux:pass` and `nsubj:pass` are all recommended by UD for passive sentence structures where the patient of a verb is expressed syntactically as the subject.

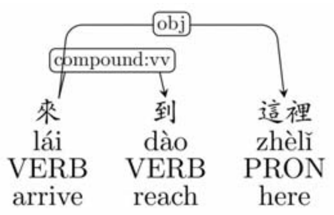
(14)



'I have been hit'

In Chinese, objects can express the goal of a displacement verb (e.g. 去 *qù* 'to go', 回 *huí* 'to return' and 來 *lái* 'to come'), such as in 我想去台灣 *wǒ xiǎng qù <u>táiwān</u>* 'I want to go to Taiwan'. When the argument of a placement verb is not introduced by a preposition, it is also annotated with `obj`, such as in: 我想住台灣 *wǒ xiǎng zhù <u>táiwān</u>* 'I want to live in Taiwan'. When the preposition is incorporated inside the verb, an alternation attested by the possibility to insert the suffix 了 *le* after it, the relation `obj` is also used: 張三跑向了電梯 *zhāngsān pǎo<u>xiàng</u>* (*le*) *<u>diàntī</u>* 'Zhangsan ran to the elevator'. The morpheme 到 *dào* 'reach' when appearing just after a movement verb is always treated as a verb forming with the movement verb a compound. Accordingly, the noun phrase introduced by it is also annotated with `obj` (15).

(15)



'arrived here.'

Treating similar phenomena at the syntactic level consistently (patient and locative objects) meets the UD criterion 5 that "UD must be suitable for computer parsing with high accuracy".

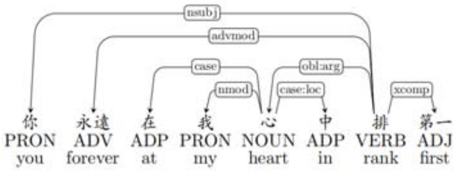
The double object construction in Chinese is fairly similar to that in English. Thus, some verbs in Chinese can take two dependents that both are coded as objects. Typically, these constructions express a transfer operation. The noun phrase expressing the recipient is an indirect object (annotated `iobj`) and the noun phrase expressing the theme is a direct object (annotated `obj`): 給他兩本書 *gěi tā liǎng běn shū* ‘Give him two books’. When the direct object is elided or is dislocated, then the noun phrase expressing the recipient is analyzed as `obj`: 今年給他吧 *jīnnián gěi tā ba* ‘This year let’s give (it to) him’. A nominal that functions as a non-core argument or a modifier is considered as oblique, and is annotated as `obl`. This relation is discussed in the next subsection. For sake of exhaustiveness, we extended further our discussion to Chinese modifiers in general, covering both nominal and adverbial modifiers.

5.3 Obliques

5.3.1 Oblique arguments and modifiers

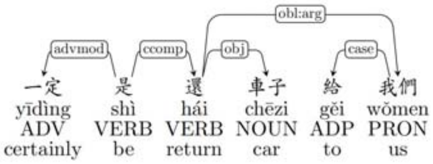
Nominal dependents of a verb that are not core arguments are obliques. Contrary to core arguments, obliques in Chinese are generally introduced or followed by an adposition. As for core arguments, word order in Chinese permits to distinguish between arguments and modifiers among obliques. In Chinese, the word order of oblique arguments is relatively free, but oblique modifiers can only appear in pre-verbal position. Oblique arguments of placement and ditransitive verbs can occupy a pre-verbal and post-verbal position (cf. 16a and 16b). On the other hand, oblique modifiers expressing location, source, and orientation can only occupy a pre-verbal position (17).

(16a)



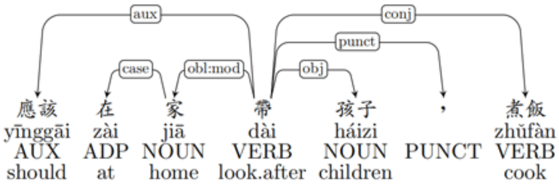
‘In my mind, you always come first!’

(16b)



‘She must be returning the car to us!’

(17)

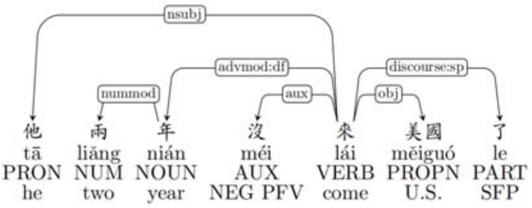


‘Should be at home looking after the children, cooking’

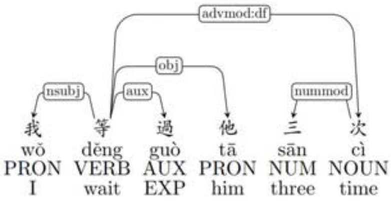
This is underlined by Paul (2015: 20): “Only arguments subcategorized for by the verb and ‘quasi’ arguments depending on the verb’s aktionsart, i.e. quantifier phrases indicating duration or frequency are admitted in postverbal position. (…) Unlike arguments, adverbs and phrasal adjuncts are totally excluded from the postverbal position in modern Mandarin.”[6] Duration and frequency adverbial phrases usually consist of a numeral and a classifier, and have the ability to appear in both preverbal (18) and postverbal (19) positions. To differentiate these quantifier phrases from nominal phrases and simple adverbs, they are given the label `advmod:df`.

[6] This excludes the cases of preposition incorporation of postverbal locative prepositional phrases mentioned in section 5.2 (cf. Peck & Lin 2019).

(18)



(19)



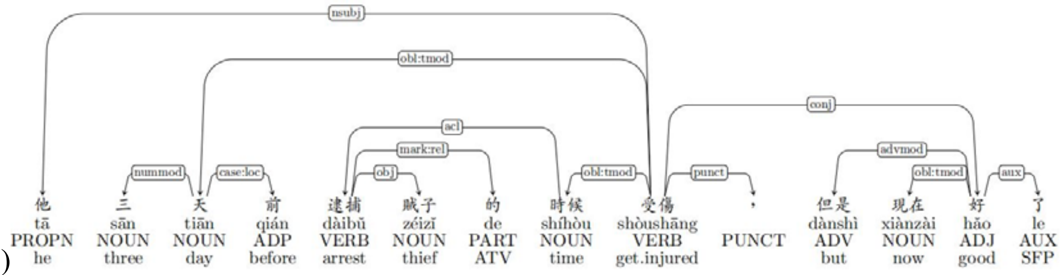
‘It has been two years that he has not come to the U.S.’ ‘I waited for him three times.’

Because UD distinguishes between nominal and clausal dependent, clausal modifiers are annotated `advcl` (which stands for adverbial clause modifier). Clausal modifiers include temporal, consequence, conditional, and purpose clauses. An illustration of a temporal clausal modifier is: 她去東京後我再加入 <u>*tā qù dōngjīng zhīhòu*</u> *wǒ zài jiārù* ‘I’ll join her after she has arrived in Tokyo’. In this section, we proposed to distinguish between argument and modifier obliques in Chinese. This improvement has not already been implemented nor evaluated in our current treebank, which only contains the relations `obl` (encompassing argument and modifier obliques) and `obl:tmod` (that should eventually be subsumed by the relation `obl:mod`). Below we discuss other usages we made of the relation `obl`.

5.3.2 Other usages of the oblique relation

Temporal nouns acting as temporal modifiers of clauses (e.g. 今天 *jīntiān* ‘today’; 去年 *qùnián* ‘last year’; 晚上 *wǎnshàng* ‘night’) are analyzed with the subrelation `obl:tmod`. This relation is employed in many UD treebanks (e.g. Arabic, Cantonese, Classical Chinese, Danish, English). Consequently, using it increases comparatibility between our Chinese treebank and other UD treebanks. Temporal nouns introduced by a preposition are also analyzed as `obl:tmod` (20).

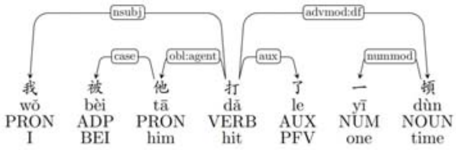
(20)



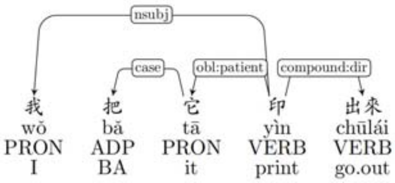
‘Three days ago, he was injured when he arrested a thief, but he has recovered now.’

As mentioned already in section 4, subsections 4.2, and 5.2, valence markers 被 *bèi* and 把 *bǎ* are treated as prepositions. When the agent of the passive construction is expressed, the preposition phrase it forms with 被 *bèi* is linked to the verb with the relation `obl:agent`, see (21). This relation is used for agents of passive constructions in many other languages (e.g. Ancient Greek, Armenian, Belarusian, Breton). Objects marked and fronted by 把 *bǎ* take the main verb as their governor, while the objects themselves serve as the governor of 把 *bǎ*, see (22). The relation employed is `obl:patient`.

(21)



(22)



‘I have been hit by him.’ ‘I print it out.’

To synthesize, the relations `nsubj`, `obj`, `iobj` and `obl` are used to annotate nominals that are core arguments and non-core arguments or modifiers of the verb. For sake of consistency with policies followed by other language treebanks in UD, we extended the use of `obl` to label temporal noun phrases and valency-alternation constructions. In the next section, we discuss the relation `compound` we used to describe patterns that show a tendency toward lexicalization, and to describe a construction specific to Chinese, i.e. the extent construction introduced by 得 *dé*.

6 Chinese compounds and the extent construction

In this section, we first discuss verb-object compounds, and then verb-verb compounds. Verb-verb compounds are subcategorized into two groups, namely result and phase compounds annotated `compound:vv`, and directional verb-verb compounds annotated `compound:dir`. The last subsection treats in depth the Chinese extent construction.

6.1 Verb-object compounds

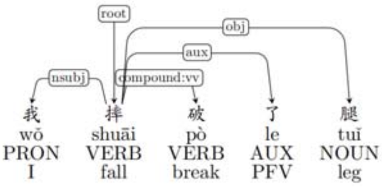
Verb-object compounds in Chinese contain a verb followed by a noun which can be separated by other words. Without intervening linguistic units, the combination though represents a single lexical unit. This aspect determines the analysis presented in this section. The verbs 睡覺 *shuìjiào* 'sleep' and 打針 *dǎzhēn* 'give/get an injection' are considered lexical verbs on their own, and represent the opposite ends of a continuum going from tight compounds to freer compounds. The former one is composed of the morphemes 睡 *shuì*, which can be used singly as a verb meaning 'to sleep' and 覺 *jiào* which functions here as a noun, but historically also meant 'to sleep'. The latter is composed of the light verb 打 *dǎ* 'hit' and the noun 針 *zhēn* 'needle'. In both cases, the noun component can still behave like a direct object of the verb component. Indeed, several syntactic and lexical behaviors show that the compound is not a tight lexical unit. Note that the (i), (ii), (iii) and (iv) behaviors are shared by the two types of verbs together, while (v) and (vi) are only verified by the type of 打針 *dǎzhēn*. It means that the latter type is less lexicalized than the former. (i) There cannot be another direct object after the compound; (ii) Aspect markers still attach directly after the verb and before the noun, such as the perfective 了 *le* in 打了幾次針 *dǎ le jǐ cì zhēn* 'Have gotten/given an injection a few times' and in 我睡了幾次覺 *wǒ shuì le jǐ cì jiào*; (iii) Duration and frequency adverbial phrases can also come in between, such as 幾次 *jǐ cì* 'a few times' in previous examples; (iv) The verb can further form a compound itself, such as in 你打完球之後 *nǐ dǎ wán qiú zhīhòu* 'After you finish playing ball' and in 我睡完覺之後 *wǒ shuì wán jiào* where the verbs 打 *dǎ* and 睡 *shuì* form compounds with the verb 完 *wán* 'finish' (see next section 6.2 for verb-verb compounds); (v) Like any direct object, the noun component can be fronted to topic position, illustrated in 你的針我還沒打 *nǐ de zhēn wǒ hái méi dǎ* 'Your injection, I still have not done it'; (vi) The noun component can be modified, shown also in the previous example where the noun 針 *zhèn* 'phone' is possessed.

Having a dedicated relation — `compound:vo` — to connect the noun component as a dependent of the verb component in these verb-object compounds acknowledges that while the noun component behaves like a direct object, the combination represents a lexical unit. To preserve this relation when the noun component is fronted (see in (v) above), we also propose `dislocated:vo` to indicate both that it is part of a compound and at the same time has been dislocated from its canonical post-verbal position. We follow the Penn Chinese word segmentation guidelines in keeping the verb and object components together as one token when there is nothing intervening in between. Therefore, we apply the dependency relation `compound:vo` only when they are separated by intervening material (see (ii)-(vi) above). The special status of these syntactically composed lexical units is not acknowledged under the Stanford Chinese and CDT annotation schemes, where the noun component is treated like any other object of a transitive verb. A researcher who wishes to extract a list of these compounds from a corpus therefore cannot do so if the corpus is annotated in either of these two schemes, but will be able to do so if it is in Chinese UD.

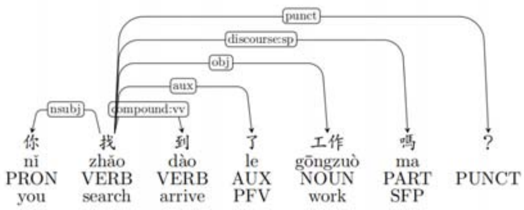
6.2 Verb-verb compounds

6.2.1 Resultative and phase compounds

We grouped together two different types of verb-verb compounds. In the first type, the second verb describes the resulting state or outcome of either the object or the subject brought on by the action of the first verb. In the second type, the second verb provides the 'phase' or (mostly telic) aspect for the first verb. The two types behave similarly in terms of their syntactic distribution, but do have some small differences which will be explained below. For the resultative compounds, both verbs may be transitive or intransitive. The second verb can also be a predicate adjective. The resulting compound is though typically causative. For example, the verbs 摔 *shuāi* 'fall' and 破 *pò* 'break' may combine, as in the example in (23), where the falling action results in the breaking of the leg. Any combination of verbs or predicate adjectives is practically possible barring semantic incompatibility (for example, you cannot *burn* something causing it to become *wet*). For the 'phase' compounds, according to Chao (1968), the second verbs may include, at the very least: the non-neutral tone versions of 着 *zháo* 'touched, got at, successful after an attempt', 到 *dào* 'arrive, reach', 見 *jiàn* 'see', 完 *wán* 'be complete, be finished', and 過 *guò* 'pass, cross' (see Chao 1968:446-450 for details). The second verb expresses the "phase of an action in the first verb" (p. 446). In this regard, they encode aspect or telicity. However, they have not fully grammaticalized to the point of tone neutralization and therefore are not quite on a par with aspect markers. The best evidence for these being different from aspect particles is that phase compounds can take aspect particles additionally, as seen in (24).

(23) 

'I fell breaking my leg.'

(24) 

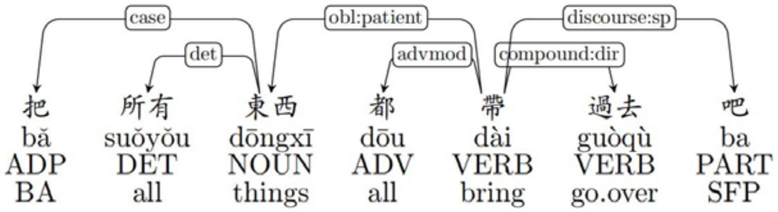
'Have you found a job?'

We propose to group the resultative and phase compounds together under the subrelation `compound:vv` because of the similarity in their syntactic behavior. (i) Aspect markers always attach after the second verb, as already seen with the aspect marker 了 *le* in (23) and (24) above; (ii) Both types of compounds can be separated by the affirmative potential 得 *de* (摔得破 *shuāi de pò* 'can break by falling'; 找得到 *zhǎo de dào* 'can find') or the negative potential 不 *bú* (摔不破 *shuāi bú pò* 'Cannot break by falling'; 找不到 *zhǎo bú dào* 'cannot find'). These potential markers alter the modality of the compound such that it expresses whether the action/state expressed by the second verb can or cannot be accomplished as a result of the action expressed by the first verb. They are tagged `PART` and linked to the second verb by the relation `mark`. Although the number of possible verb-verb combinations is practically unrestricted, these verb-verb combinations can only take one aspect marker after the entire compound and can also only take one direct object at most, making them behave like a single syntactic verb. The equivalent relations in Stanford Chinese and CDT are `rcomp` 'resultative complement' and `CMP` 'complement', respectively. In both these schemes, the equivalent relations also cover directional verb compounds, addressed bellow.

6.2.2 Directional compounds

The directional verb compound consists of a second verb which may be a single verb expressing deictic motion, specifically the two verbs 來 *lái* ‘come’ and 去 *qù* ‘go’, or other motion verbs limited to 上 *shàng* ‘go up’, 下 *xià* ‘go down’, 出 *chū* ‘exit’, 進 *jìn* ‘enter’, 回 *huí* ‘return’, 過 *guò* ‘cross’, 開 *kāi* ‘open’, and 起 *qǐ* ‘rise’. The latter motion verbs can also combine with the deictic motion verbs, such as 回來 *huílái* ‘come back’, 上來 *shànglái* ‘come up’, 下去 *xiàqù* ‘go down’, 進去 *jìnqù* ‘go in’, etc. Although all of the above directional verbs (individually or combined) can function as main verbs on their own, they can also follow a main verb adding directional and deictic information to the main verb. For example, in (25), the main verb 帶 *dài* ‘bring’ combined with the directional verb 過去 *guòqù* ‘go over’ conveys the action is away from the speaker and/or interlocutor and involves crossing over some distance.

(25)



‘How about you take all the stuff over there?’

Directional verb compounds can have idiomatic meanings that make them similar to phrasal verbs in some European languages, where the directional verb (or preposition/verbal particle such as in English) no longer refers to spatial direction, such as in 我想不出來 *wǒ xiǎng bù chūlái* ‘I can’t come up with (something)’. Besides the fact that the affirmative and negative potential 得 *de* and 不 *bù* can intervene between the first verb and the directional verb just like in the previous example, the directional verb compound can also be intervened by a direct object, as seen in (26) below. A unique characteristic of directional verb compounds is that the combined directional verbs can be intervened by a noun indicating a location, as illustrated in (27) with the noun 山 *shān* ‘mountain’.

(26)

帶	他們	出去
dài	*tāmen*	*chūqù*
take	them	go.out

‘take them out’

(27)

爬	上	山	去
pá	*shàng*	*shān*	*qù*
climb	up	mountain	go

‘climb up the mountain (away from speaker)’

Treating these directional verbs as separate from the main verb would not make sense since they do not introduce new events occurring separately from the main verb. On the other hand, these compounds should be differentiated from verb-verb compounds for several reasons. First, direct objects can come in between the main verb and the directional verb (or in intervening between the combined directional verbs), while in verb-verb compounds the direct object must occur after the two verbs. Second, the directional verbs which can form a directional compound with a main verb are a closed set, and they are semantically dedicated in terms of conveying directional and deictic information. We therefore propose the subrelation `compound:dir` specifically for these multi-word constructions. Separating these directional compounds from the resultative and phase compounds labeled with `compound:vv` is motivated not only on the syntactic differences mentioned above, but also on easier comparison and correspondence with languages which supplement directional information to verbs with adpositional particles. Both Stanford Chinese and CDT conflate the two structures with one label, `rcomp` (‘resultative complement’) and `CMP` (‘complement’) respectively. However, it is actually easy — and we think useful for cross-linguistic comparison — to distinguish them on both semantic and syntactic grounds. By distinguishing between resultative/phase compounds and directional compounds we meet the UD criterion 6: "UD must support well downstream understanding tasks". Note that these verb-verb compounds, also known as serial verb constructions (SVC), are very common in many African and South-Asian languages (Bisang 2009; Haspelmath 2016). In the Amharic, Armenian, Marathi, Naija, Telugu, Wolof and Yoruba UD treebanks they are annotated as `compound:svc`. That is, our annotation shares with these other language annotations the main relation `compound`, but not the subrelation `:svc`. Our choice is thus a trade-off between criteria 2 and 6. Finally, verb-verb compounds must be distinguished from subject and object control constructions, annotated with `xcomp`. In control constructions (e.g. 我打算去 *wǒ dǎsuàn qù* 'I plan to go'), a verb takes another one as object, and the latter lacks an overt subject. Additionally, these constructions do not denote result, phase or direction, and no potential marker can intervene between the two verbs.

Potentially, tighter patterns (like in 吃不完 *chībùwán* 'not being able to finish') could be annotated with `compound`, while freer patterns (like in 打了幾次針 *dǎ le jǐ cì zhēn* 'have gotten/given an injection a few times') could be annotated with `advcl`. But by treating all verb-object and verb-verb patterns as compounds rather than taking into account the lexicon-syntax continuum, we also meet the UD criterion 3: "UD must be suitable for rapid, consistent annotation by a human annotator." We have shown how we classify them into three subrelations according to the second element:

`compound`

↳`compound:vo`: verb-object compounds

↳`compound:vv`: resultative and phase verb-verb compounds

↳`compound:dir`: directional verb-verb compounds

The fine distinction is based on the syntactic properties of each type. However, these distinctions may not be easily understood by and useful to non-linguists. It might be enough for them only to be aware of the big category `compound`. Therefore, for the sake of coarse analysis by non-linguists, the upper-level relation `compound` can be used, while for fine analysis by linguists, the sub-relations can be used. We thus meet criterion 4: "UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing."

6.3 Extent constructions

In a particular set of Chinese constructions, a clause ending with a verb or a predicate adjective is followed by the morpheme 得 *dé* and is subsequently followed by another clause containing just a predicate adjective, a verb, or a full clause. If the second clause contains a predicate adjective, it semantically behaves like an adjective or an adverb describing the action in the first clause, as illustrated in (28). On the other hand, if the second clause is a verb, a verb phrase, or a full clause, it describes a state of affairs that is an extension or a result of the first clause, as seen in (29) where the second clause contains a verb phrase without a subject (想吐了 *xiǎng tù le* 'want to vomit') and in (30) where it is a full clause.

(28)

你	說	得	對
nǐ	*shuō*	*dé*	*duì*
you	say	DE	correct

'you said it correctly'

(29)

我	熱	得	想	吐	了。
wǒ	*rè*	*dé*	*xiǎng*	*tǔ*	*le*
I	hot	DE	want	vomit	SP

'I'm so hot I want to vomit.'

(30)

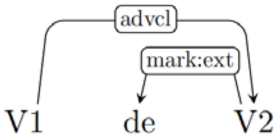
熱	得	連	塑膠袋	也	出汗。
rè	*dé*	*lián*	*sùjiāodài*	*yě*	*chūhàn*

hot DE even plastic.bag also sweat
‘It’s so hot even plastic bags are sweating.’

In the literature, the first type containing the predicate adjective is traditionally called a “depictive” or “descriptive complement construction”. The second type with a verb is called a “resultative complement construction” (Huang 1988). To avoid overlap with other Chinese constructions bearing similar names, and to unify the two mentioned related constructions, we refer to them as “extent constructions” for short, after Chao (1968). For the purpose of analytical exposition in this section, we will schematize all of these cases as [V1 *de* V2] where V stands for either a verb or a predicate adjective, and both V1 and V2 may have additional arguments or modifiers, with the major caveat that no matter what other additional elements are present, 得 *dé* must always come immediately after V1.

Given that there are two predicates in these constructions, a natural question is whether the first or the second predicate is the main predicate. This question has been quite hotly debated in the literature and propositions have been made in favor of both analyses (Huang 1988). Newer research has continued this debate — including Zhang (2001), Huang et al. (2009), Chen (2012), and Li (2015) which support the analysis that the V1-predicate is the main predicate, and Wei (2006) and Osborne and Ma (2015) which support the analysis that at least one if not both types of the construction have the V2-predicate as the main predicate. We recognize that the debate on the true syntactic nature of extent constructions is still ongoing, so while we propose adopting the annotation strategy illustrated in (31). We think it would be more appropriate to give a language-specific relation to this unique construction in Chinese. This way, it may be easily converted if future research show that a different annotation strategy is preferable. We propose to link V1 to V2 with `advcl`, and 得 *dé* to V1 with `mark:ext`. Finally, the morpheme 得 *dé* is tagged `PART`.

(31)



We were unable to find explicit guidelines or examples of how this structure should be annotated in Stanford Chinese and CDT, and believe this may be the first explicit treatment of this structure in a dependency annotation scheme for Chinese. By clarifying the extent constructions analysis and by distinguishing Chinese compounds, we meet the UD criterion 1 “UD needs to be satisfactory on linguistic analysis grounds for individual languages”.

In these sections, we showed how we applied the UD scheme to Mandarin Chinese. Below we present the resource annotated with this scheme, and the evaluation of its annotation.

7 The UD Chinese HK treebank

7.1 Content and comparison

In this section, we first describe the content of the UD Chinese HK treebank, with a comparison to the other existing UD treebanks for Chinese. Then we report on the annotation procedure used for our evaluation, and on the inter-annotator agreement. At the time this paper is written, UD v2.8 has been released, and there are five Chinese UD treebanks. They can be classified according to the research team who mainly created them. Annotation choices in each treebank vary from one team to the other.

- Google
    - UD Chinese GSD. Traditional Chinese Universal Dependencies treebank annotated and converted by Google. Contains 123,291 tokens and 4,997 sentences.

  - UD Chinese GSDSimp. Simplified Chinese Universal Dependencies treebank converted from the GSD (traditional) treebank with manual corrections. Contains 123,291 tokens and 4,997 sentences.
  - UD Chinese PUD. Traditional Chinese part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from raw text to Universal Dependencies. Contains 21,415 tokens and 1,000 sentences.
- City University of Hong Kong
  - UD Chinese CFL. Simplified Chinese essays written by learners of Chinese as a foreign language, annotated by researchers at the City University of Hong Kong. Contains 7,256 tokens and 451 sentences.
  - UD Chinese HK. Traditional Chinese part of the Mandarin Chinese-Cantonese parallel treebank annotated by researchers at the City University of Hong Kong. Contains 9,874 tokens and 1,004 sentences.

The guidelines presented here have been developed during and for the construction of the UD Chinese HK treebank. The UD Chinese HK treebank is composed of two sub-corpora. One, which represents the colloquial register, is made of the Chinese subtitles of three Cantonese short movies that were produced by students from the creative media programme at City University of Hong Kong. The other, which represents the formal register, is made of part of the proceedings of the Hong Kong S.A.R. Legislative Council 12th October 2016 meeting. This treebank belongs to the endeavor of creating a Cantonese-Mandarin parallel corpus for the contrastive study of the syntax of these two languages (Lee 2011; Wong et al. 2017). To fulfill this purpose, we have chosen Cantonese audio-visual materials with Mandarin translation, rather than original Mandarin Chinese texts. The UD Chinese HK treebank is available publicly (License CC BY-SA 4.0), and is posted on the UD portal. The size of the corpus in terms of number of sentences and tokens are given in Table 3 below, where the figures exclude punctuation.

Table 3. Statistics on the Mandarin UD Chinese HK treebank

	Short Movies	**Legislative Council Proceedings**	**Total**
Sentences	650	354	1004
Tokens	3,824	4,312	8,136
Average sentence length	5.88	12.18	8.1

It can be seen in Table 3 that the sentences are relatively short — 5.88 tokens on average — in the Short Movies sub-corpus, while longer sentences — 12.18 tokens on average — are observed in the Legislative Council Proceedings sub-corpus. The distribution of the sentence length in the UD Chinese HK treebank is shown in Figure 1 below. The Short Movies sub-corpus has most of its sentences (88.6%) in between 1 and 10 tokens of length, while the Legislative Council Proceedings sub-corpus has a more homogeneous distribution, with most of the sentences (88.7%) being situated in between 1 and 20 tokens of length.

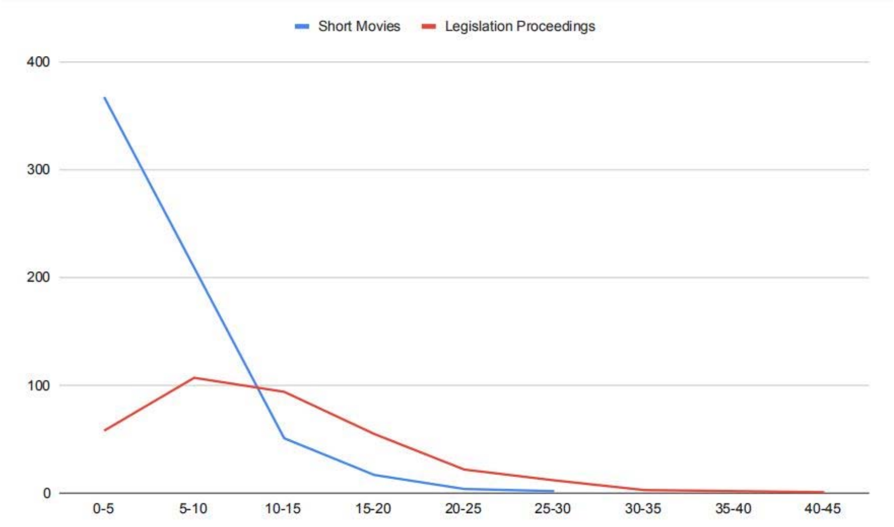


Figure 1. Distribution of sentences length in the UD Chinese HK treebank

As noted earlier, the annotation choices vary from one treebank to the other, according to the team behind their creation. The CFL treebank created at the City University of Hong Kong follows the same guidelines presented in this paper. Taking the perspective of this pair of treebank, the `advmod:df`, `compound:dir`, `compound:vo`, `compound:vv`, `mark:rel` are not used in the other three treebanks. The `case:loc`, `discourse:sp`, `mark:adv`, `obl:agent`, `obl:tmod` have been adopted by PUD, but not by GSD and GSDSimp. Conversely, the subrelations `aux:aspect`, `case:dec`, `case:pref`, `case:suff`, `csubj:pass`, `flat:foreign`, `mark:advb`, `mark:comp`, `mark:relcl`, `nmod:tmod` are used in GSD and GSDSimp, but not in the three others. Without going more deeply into details, one treatment of the valence markers 把 *bǎ* and 被 *bèi* is shared by HK/CFL and PUD, but not by GSD and GSDSimp, and one treatment of classifiers is shared by GSD, GSDSimp and PUD, but not by HK/CFL. To convert the relations from other treebanks annotation scheme to the UD Chinese HK scheme, the mapping rules might be rather complicated, while that of the subrelations might be more straightforward.

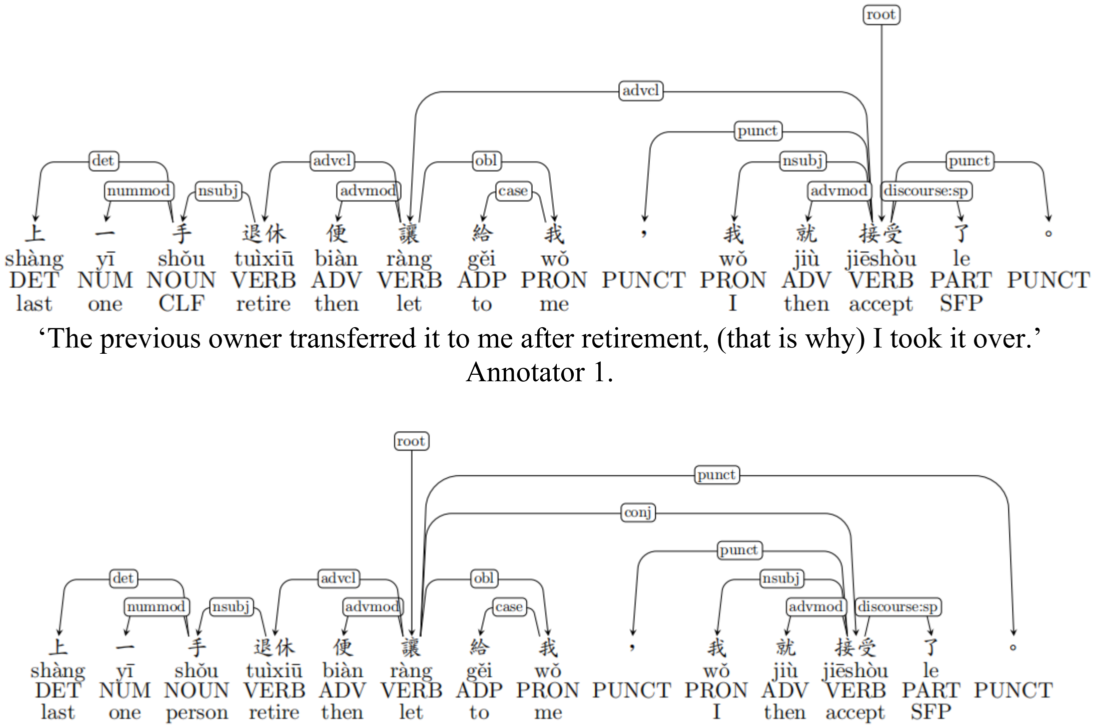
7.2 Evaluation

For the evaluation, two annotators independently annotated 109 Chinese sentences from the subtitles of a Cantonese short movie. We organized three rounds of agreement annotation. In Round 1, annotators separately annotated their own sentences. In Round 2, both annotators were asked to go over the annotation guidelines again, and were also given access to each other's trees. They were given a chance to review and change their annotations. In Round 3, the annotators were asked to discuss their differences from Round 2, and agree on the most accurate annotation where possible, or otherwise note down those sentences for which they think that more than one analysis are valid, while retaining the annotation judged the most appropriate for the context. These cases counted as disagreements between the annotators. The agreement figures are summarized in Table 4. Following Berzak et al. (2016), the agreement is measured as the fraction of agreed labels between the two annotators, the Cohen's Kappa scores (Cohen 1960) for POS tags and dependency labels are also provided. Note that the evaluation sample size does not include punctuation.

Table 4. Inter-annotator agreement

	Round 1	**Round 2**	**Round 3**
Number of sentences	109		
Number of tokens	1077		
Tokens per sentence	9.88		
Tokenization	Precision: 0.9959 Recall: 0.995	--	--
POS	0.9415	0.9786	0.9991
POS Cohen’s kappa	κ = 0.9313	κ = 0.975	κ = 0.9989
Heads	0.7976	0.8394	0.9712
Relations	0.779	0.8644	0.9582
Relations Cohen’s kappa	κ = 0.8764	κ = 0.9546	κ = 0.9722
Heads + Relations	0.7066	0.8041	0.9461
POS + Heads + Relations	0.6834	0.7985	0.9452

As can be seen from Table 4, the area of disagreement concerns mostly dependency choices, i.e. the choice of the relation and the choice of the head. In Chinese, conjunction words can be and often are omitted, such that two clauses are simply juxtaposed next to each other. This leaves the relationship between the two clauses to the subjective interpretation of each annotator. The two most frequent types of disagreements are related to this peculiarity. First, the two annotators often disagreed on whether the two clauses are coordinated or run-on sentences. In the first case, the relation is `conj`, in the second it is `parataxis`. Second, they disagreed on whether the two clauses are coordinated or whether one is the subordinated clause of the other. If two clauses are identified as coordinated by the annotator, the relation is usually `conj`. For subordinated clauses, the relation is usually `advcl`. A disagreement between the two annotators on such a case is illustrated in Figure 2.

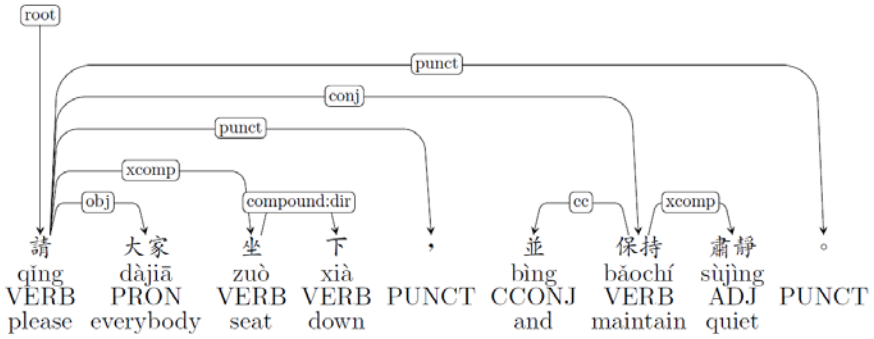


‘The previous owner transferred it to me after retirement, (that is why) I took it over.’
Annotator 1.

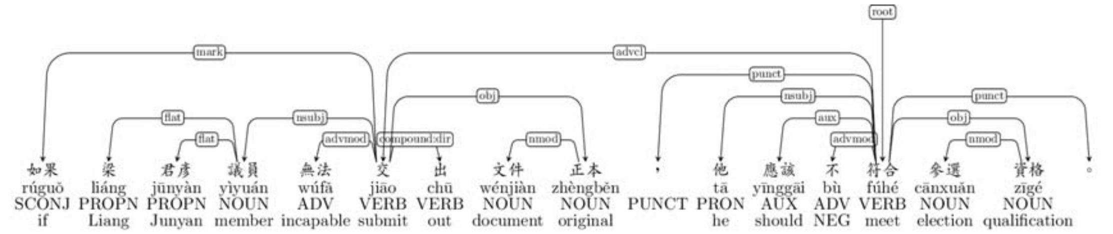
‘The previous owner transferred it to me after retirement; then I took it over.’
Annotator 2.

Figure 2. Disagreement on relations

The sentences in Figure 3 from the Legislative Council Proceedings part of the corpus illustrate the case where the structural relation between two clauses is made explicit by conjunction words.



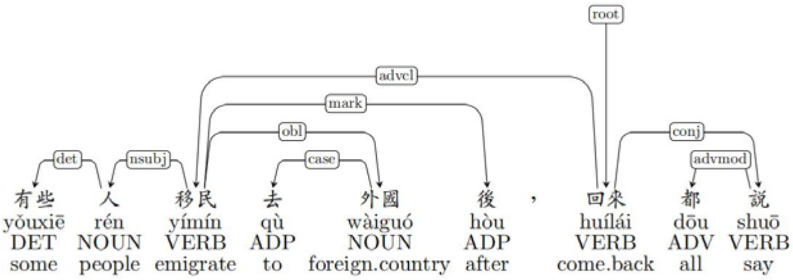
‘Please sit down and keep quiet’



‘If Mr. Junyan Liang cannot produce the original document,
he should not be eligible to run for the election.’

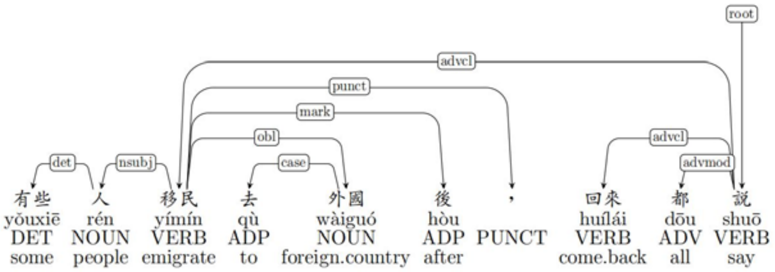
Figure 3. The marking of the structural relation between two clauses

As for disagreement concerning the head, there were 13 sentences in which the annotators disagreed. In those cases, more than one valid analysis is possible. These also mostly deal with sentences containing at least two clauses in which it was possible to argue that either one is the main clause. This situation is illustrated in Figure 4.



Annotator 1.

‘After some people emigrate to other countries, they come back and all say…’



Annotator 2.

‘After some people emigrate to other countries, and after they come back, they all say…’

Figure 4. Disagreement on heads

In Figure 4, the sentence 有些人移民去外國後，回來都說 *yǒuxiē rén yímín qù guó wài hòu, huílái dōu shuō* can have two different interpretations. The first interpretation (Annotator 1) is ‘After some people emigrate to other countries, they come back and all say…’. It triggers a parsing with the verb 回來 *huílái* ‘come back’ as the head of the second clause. The second interpretation (Annotator 2) is ‘After some people emigrate to other countries, and after they come back, they all say…’. It triggers an analysis of the verb 說 *shuō* ‘say’ as the head of the second clause. Despite these areas of inconsistencies in annotation, the inter-annotator agreement still reached a satisfactory level. This measure quantitatively attests that our proposed UD scheme for Chinese fulfills the criterion 3: “UD must be suitable for rapid, consistent annotation by a human annotator”, while it has been designed with careful consideration of all other criteria, as we have made clear in the different sections.

8 Conclusion and future work

We have presented a Universal Dependencies (UD) scheme, for the creation of the UD Chinese HK treebank, with consideration and reference to two other dependency schemes previously created for this language. The morphosyntactic treatment of classifiers as nouns was motivated, and explicit criteria for the definition of Chinese adpositions were provided. We presented how we have extended or adapted the definition of five relations, and proposed or adapted eleven new subrelations. We have motivated our design with reference to Chinese linguistic analysis and best practices from existing dependency schemes for Chinese.

In line with UD criteria, the proposed scheme can be expected to lead to more accurate analysis of Chinese (e.g., Chinese compounds, extent constructions in Section 6, Criterion 1); promote cross-linguistic parallelism across language families (e.g. classifiers in Section 4.1, Criterion 2); improve annotation consistency (e.g., Chinese compounds; extent constructions in Section 6, Criterion 3); enable more accurate automatic parsing (e.g., patient and locative objects in Section 5.2, Criterion 5); facilitate comprehension by non-linguists (e.g., compound relation in Section 6, Criterion 4); and support downstream understanding tasks (e.g., resultative/phase verb-verb compounds and directional verb-verb compounds and in Section 6, Criterion 6).

This annotation scheme has been applied to our corpus. In an evaluation of this corpus annotation, we have shown that our proposed UD scheme for Chinese also ensures high degree of agreement between annotators, while it is motivated by sound linguistic analysis, and lends it self to cross-linguistic comparison,

Once a large enough set of sentences is manually annotated to form training and test sets, it would be interesting to use a real state-of-the-art parser to verify that our annotation scheme is suitable for high accuracy parsing. The mapping from previous treebanks such as the CTB and CDT based on our correspondence charts, as well as the homogenization of the different UD Chinese treebanks based on the guidelines presented in this paper should be carried out in future works.

Glosses

ATV	attributive and possessive particle 的 *de*
BA	valence marker 把 *bǎ*
BEI	valence marker 被 *bèi*
CLF	classifier
DE	extent marker 得 *dé*
EXP	experiential aspect marker 過 *guò*
NEG	negative adverb 不 *bù*

NEG_PFV	negative perfective aspect marker 沒 *méi*
PFV	perfective aspect marker 了 *le*
SFP	sentence particle 了 *le*

References

Aikhenvald, A. Y. (2000). *Classifiers: A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.

Andrews, D. A. (2007). The Major Functions of the Noun Phrase. In: Shopen, T. (Ed.). *Language Typology and Syntactic Description*, Vol. 1. (pp. 132-133). Cambridge: Cambridge University Press.

Berzak, Y., Kenney, J., Spadine, C., Wang, J., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016). Universal Dependencies for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 737-746.

Bisang, W. (2009). Serial Verb Constructions. *Language and Linguistics Compass*, *3*(3), 792-814.

Chang, P. (2009). Improving Chinese-English Machine Translation through Better Source-side Linguistic Processing. Ph.D. Dissertation, Stanford University.

Chang, P., Tseng, H., Jurafsky, D., & Manning, D. C. (2009). Discriminative Reordering with Chinese Grammatical Relations Features. In: *Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation*. Boulder, United States, pp. 51-59.

Chao, Y. (1968). *A Grammar of Spoken Chinese.* Berkeley & Los Angeles: University of California Press.

Che, W., Li, Z., & Liu, T. (2012). Chinese Dependency Treebank 1.0 LDC2012T05. Linguistic Data Consortium, Philadelphia. https://catalog.ldc.upenn.edu/LDC2012T05

Chen, Y. (2012). *Hao-De-Xia-Ren*: The Grammatical Property of *De* in V-*De*-C Construction. *Kaohsiung Normal University Journal* [高雄師大學報], *32*, 75-98.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1): 37-47.

Croft, W., Nordquist, D., Looney, K., & Regan, M. (2017). Linguistic Typology Meets Universal Dependencies. In: *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*. Bloomington, United States, pp. 63-75.

Djamouri, R., Paul, W., & Whitman, J. (2013). Postpositions vs Prepositions in Mandarin Chinese: The Articulation of Disharmony. In: Biberauer, B. and Sheehan, M. (Eds.), *Theoretical Approaches to Disharmonic Word Order* (pp. 4-105). Oxford: Oxford University Press.

Gerdes, K., & Kahane, S. (2016). Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In: *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016*. Berlin, Germany, pp. 131-140.

Haspelmath, M. (2016). The Serial Verb Construction: Comparative Concept and Cross-Linguistic Generalizations. *Language and Linguistics*, *17*(3), 291-319.

Harbin Institute of Technology Research Center for Social Computing and Information Retrieval [HIT-SCIR] (2010). HIT-CIR Chinese Dependency Treebank Annotation Guideline (HIT-CIR 汉语依存树库标注规范).

Huang, C.-T. J. (1988). *Wo Pao De Kuai* and Chinese Phrase Structure. *Language*, *64*, 274-311.

Huang, C.-T. J. (1992). Complex Predicates in Control. In: Larson, R. et al. (Eds.), *Control and Grammar* (pp. 109-147). Dordrecht: Kluwer.

Huang, C.-T. J, Li, A., & Li, Y. (2009). *The Syntax of Chinese.* Cambridge & New York: Cambridge University Press.

Lee, J. S. (2011). Toward a Parallel Corpus of Spoken Cantonese and Written Chinese. In: *Proceedings of the 5th International Joint Conference of Natural Language Processing*. Chiang Mai, Thailand, pp. 1462-1466.

Leung, H., Poiret, R., Wong, T.-S., Chen, X., Gerdes, K., & Lee, J. S. (2016). Developing Universal Dependencies for Mandarin Chinese. In: *Proceedings of the 12th Workshop on Asian Language Resources*. Osaka, Japan, pp. 20-29.

Li, C. (2015). On the V-*DE* Construction in Mandarin Chinese. *Lingua Sinica*, *1*(6), 1-40.

Li, C., & Thompson, A. S. (1981). *Mandarin Chinese: A Functional Reference Grammar.* Berkeley & Los Angeles: University of California Press.

Lu, B., Zhang, G., & Bisang, W. (2015). Valency Classes in Mandarin. In: Malchukov, A. and Comrie, B. (Eds.), *Valency classes in the world's languages* (pp. 709-764). Berlin: Mouton de Gruyter.

de Marneffe, M.-C., MacCartney, B., & Manning., D. C. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova, Italy, pp. 449-454.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, G., Nivre, J., & Manning., D. C. (2014). Universal Stanford Dependencies: A Cross-Linguistic Typology. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavik, Iceland, pp. 4584-4592.

de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, D. C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., & Zeman, D. (2016). *Universal Dependencies v2*. http://universaldependencies.org/

de Marneffe, M.-C., Manning, D. C., Nivre J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, D. C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 1659-1666.

Nivre, J., de Marneffe, M.-C., Ginter, F. Hajič, J., Manning, D. C., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: *Proceedings of the 12th International Conference on Language Resources and Evaluation*. Marseille, France, pp. 4034-4043.

Osborne, T., & Ma., S. (2015). A DG Account of the Descriptive and Resultative *de*-Constructions in Chinese. In: *Proceedings of the 3rd International Conference on Dependency Linguistics*. Uppsala, Sweden, pp. 261-270.

Osborne, T., & Gerdes, K. (2019). The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD). *Glossa*, *4*(1).

Paul, W. (2015). *New Perspectives on Chinese Syntax.* Berlin: De Gruyter.

Peck, J., & Lin, J. (2019). Semantic Constraint on Preposition Incorporation of Postverbal Incorporation of Postverbal Locative PPs in Mandarin Chinese. *Languages and Linguistics*, *20*(1), 85-130.

Tremblay, A. (2005). Word Order in Mandarin Chinese and Grammatical Relations. In: *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation*. Taipei, Taiwan, pp. 333-340.

Tremblay, A., & Beck, D. (2013). Semantic-Communicative Structure and Word Order in Mandarin Chinese. *Open Journal of Modern Linguistics*, *3*(1), 79-86.

Wei, J. (2006). Two Types of V-*DE* Constructions in Mandarin Chinese. *UST Working Papers in Linguistics*, *2*, 97-107.

Wong, T.-S., Gerdes, K. Leung, H., & Lee, J. S. (2017). Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In: *Proceedings of the 4th International Conference on Dependency Linguistics*. Pisa, Italy, pp. 266-275.

Xia, F. (2000a). The Segmentation Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-06, http://repository.upenn.edu/ircs_reports/37/

Xia, F. (2000b). The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-07, http://repository.upenn.edu/ircs_reports/38/

Xue, N, Xia, F., Huang, S., & Kroch., A. (2000). The Bracketing Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-08, http://repository.upenn.edu/ircs_reports/39/

Xue, N., Zhang, X., Jiang, Z., Palmer, M., Xia, F., Chiou, F., & Chang, M. (2013). Chinese Treebank 9.0 LDC2016T13. Linguistic Data Consortium, Philadelphia, https://catalog.ldc.upenn.edu/LDC2016T13

Zeman, D. (2017). Core Arguments in Universal Dependencies. In: *Proceedings of the 4th International Conference on Dependency Linguistics*. Pisa, Italy, pp. 287-296.

Zhang, M., Zhang, Y., Che, W., & Liu. T. (2014). A Semantics Oriented Grammar for Chinese Treebanking. In: *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*. Kathmandu, Nepal, pp. 366-378.

Zhang, N. (2001). The Structures of Depective and Resultative Constructions in Chinese. *ZAS Papers in Linguistics*, *22*, 191-221.

Appendices

Appendix 1. Parts of speech correspondence chart

UD Chinese HK	Penn Chinese (CTB)	Harbin (CDT)
ADJ	JJ (Adjective) VA (Predicative adjective) OD (Ordinal number)	a (Adjective) m ((Number))
ADP	P (Preposition) BA (把 *bǎ* [and 將 *jiāng*] in 把 *bǎ*-construction) LB (被 *bèi* in long 被 *bèi*-construction) LC (Localizer)	p (Preposition) nd (Direction noun)
ADV	AD (Adverb)	d (Adverb)
AUX	AS (Aspect particle) SB (被 *bèi* in short 被 *bèi*-construction) VV ((Other verbs))	u ((Auxiliary)) v ((Verb))
CCONJ	CC (Coordinating conjunction)	c ((Conjunction))
DET	DT (Determiner)	r ((Pronoun)) b ((Other noun-modifier))
INTJ	IJ (Interjection)	e (Exclamation)
NOUN	M (Measure word) NN (Other noun) NT (Temporal noun)	q (Quantity) n (General noun) ni (Organization name) nl (Location noun) nt (Temporal noun)
NUM	CD ((Cardinal number))	m ((Number))
PART	DEC (得 *dé* as a complementizer/nominalizer) DEG (得 *dé* as a genitive/associative marker) DER (Resultative 得 *dé*) DEV (Manner 地 *de*) ETC ('etc.' marker 等 *děng*) SP (Sentence-final particle 嗎 *mǎ*) MSP ((Other particle))	u ((Auxiliary))
PRON	PN (Pronoun)	r ((Pronoun))

PROPN	NR (Proper noun)	nh (Person name) ns (Geographical name) nz (Other proper noun)
PUNCT	PU (Punctuation)	wp (Punctuation)
SCONJ	CS (Subordinating conjunction)	c ((Conjunction))
SYM	PU?	wp?
VERB	VC (Copula) VE (Existential/possessive verbs as main verbs) VV ((Other verb))	v ((Verb))
X	FW (Foreign word)	ws (Foreign words)
[based on context]	?	b (Other noun-modifier) i (Idiom) j (Abbreviation) o (Onomatopeia)
[treated as part of a token and/or based on context]		h (Prefix) k (Suffix)

Appendix 2. Comparison of dependency systems—standard UD relations

Chinese UD	Stanford Chinese	Harbin (CDT)
acl (Clausal modifier of noun)	rcmod (< Relative clause) vmod (< Verb modifier)	ATT ((Attribute))
advcl (Adverbial clause modifier)	--	ADV ((Adverbial))
advmod (Adverbial modifier)	advmod (< Adverbial modifier) dvpmod (< Manner 的 *de* modifier)	ADV ((Adverbial))
amod (Adjectival modifier)	amod (Adjectival modifier)	ATT ((Attribute))
appos (Appositional modifier)	prnmod ((< Parenthetical modifier))	COO ((Coordinate))
aux (Auxiliary)	mmod (Modal modifier)	RAD ((< Right adjunct))
case (Case)	assm (< Associative modifier) pobj ↺ (< Prepositional object) plmod ↓ ↺ (< Localizer modifier of a prep.)	POB ↺ ((Preposition-object))

`cc` (Coordinating conjunction)	`cc` ↑ (Coordinating conjunction)	`LAD` ↑ ((Left adunct))
`ccomp` (Clausal complement)	`ccomp` (Clausal complement)	`VOB` ((Object of verb))
`clf` (Classifier modifier)	`clf` (Classifier modifier)	`ATT` ((Attribute))
`compound` (Compound)	`nn` (< Noun compound modifier)	`ATT` ((Attribute))
`conj` (Conjunct)	`conj` ↺ (< Conjunct) `comod` (< Coordinated verb compound modifier)	`COO` ((Coordinate))
`cop` (Copula)	`cop` (< Copular) `attr` ↺ (< Attributive)	`VOB` ↺ ((Object of verb))
`csubj` (Clausal subject)	--	`SBV` ?
`det` (Determiner)	`det` (Determiner)	`ATT` ((Attribute))
`dep` (Unspecified dependency)	--	--
`discourse` (Discourse element)	--	--
`dislocated` (Dislocated element)	--	`FOB` ((Fronting-object))
`fixed` (Fixed MWE)	--	--
`flat` (Flat MWE)	`nn` ↺ ?	--
`goeswith` (Tokenization connector)	--	--
`iobj` (Indirect object)	`range` (< Dative object that is a quantifier phrase)	`IOB` (Indirect object)
`list` (List)	--	--
`mark` (Marker)	`prtmod` (< Particles such as 所以 *suǒyǐ* ‘all’; 来 *lái* ‘to’; 而 *ér* ‘and’) `lccomp` ↺ (< Clausal complement of a localizer) `pccomp` ↺ (< Clausal complement of a preposition)	`ADV` ((Adverbial)) `ATT` ((Attribute)) `LAD` ((Left adjunct))
`nmod` (Nominal modifier)	`assmod` (< Associative modifier)	`ATT` ((Attribute))
`nsubj` (Nominal subject)	`nsubj` (< Nominal subject) `top` (< Topic)	`SBV` (Subject of verb)
`nummod` (Numeric modifier)	`nummod` (< Number modifier) `ordmod` (< Ordinal	`ATT` ((Attribute))

	number modifier)	
obj (Object)	dobj (Direct object)	DBL (< Double roles: subj. & obj.) VOB ((Object of verb))
obl (Oblique nominal)	ba 把 *bǎ* ↓ (< BA) loc ↓ brep ↓ (< Coverbs and prepositions) [these can also be in the 'case' row with ↑ instead of ↓]	ADV ↓ ((Adverbial))
orphan (Orphan)	--	--
parataxis (Parataxis)	prnmod ((< Parenthetical modifier))	COO ((Coordinate)) IS (Independent structure)
punct (Punctuation)	punct (Punctuation)	WP (Punctuation)
reparandum (Overridden disfluency)	--	--
root (Root)	root (Root)	HED (Head)
vocative (Vocative)	--	--
xcomp (Open clausal complement)	xsubj ↓ ↺ (< Controlling subject)	VOB (Object of verb)

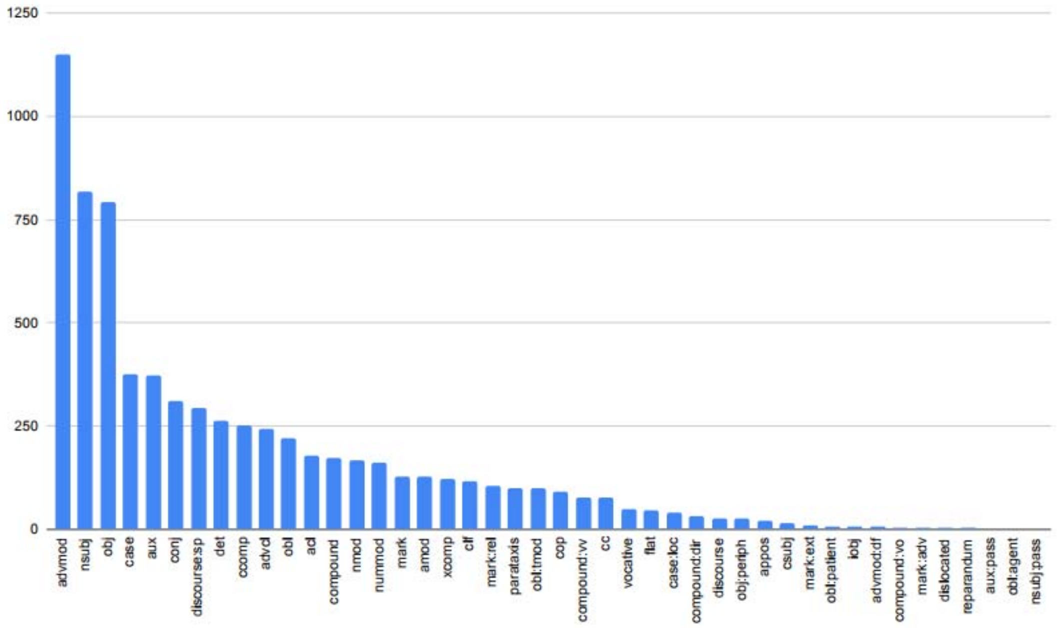
Appendix 3. Comparison of dependency systems—proposed language-specific relations in Chinese UD v2

UD Chinese HK	Stanford Chinese	Harbin (CTD)
advmod:df (Adverbial modifier: duration and frequency)	--	CMP ((Complement))
aux:pass (Passive auxiliary)	pass ((Passive marker))	ADV ((Adverbial))
case:loc (Localizer)	lobj ↺ (Localizer object)	ATT ↺ ((Attribute))
compound:dir (Directional verb compound)	rcomp ((Resultative complement))	CMP ((Complement))
compound:vo (Verb-object compound)	dobj ((Direct object)) ?	DBL (< Double roles: subject & object') ? VOB ((Object of verb)) ?
compound:vv (Verb-verb compound)	rcomp ((Resultative complement))	CMP ((Complement))
csubj:pass (Clausal passive subject)	--	FOB?
discourse:sp (Sentential particle)	--	RAD ((Right adjunct))

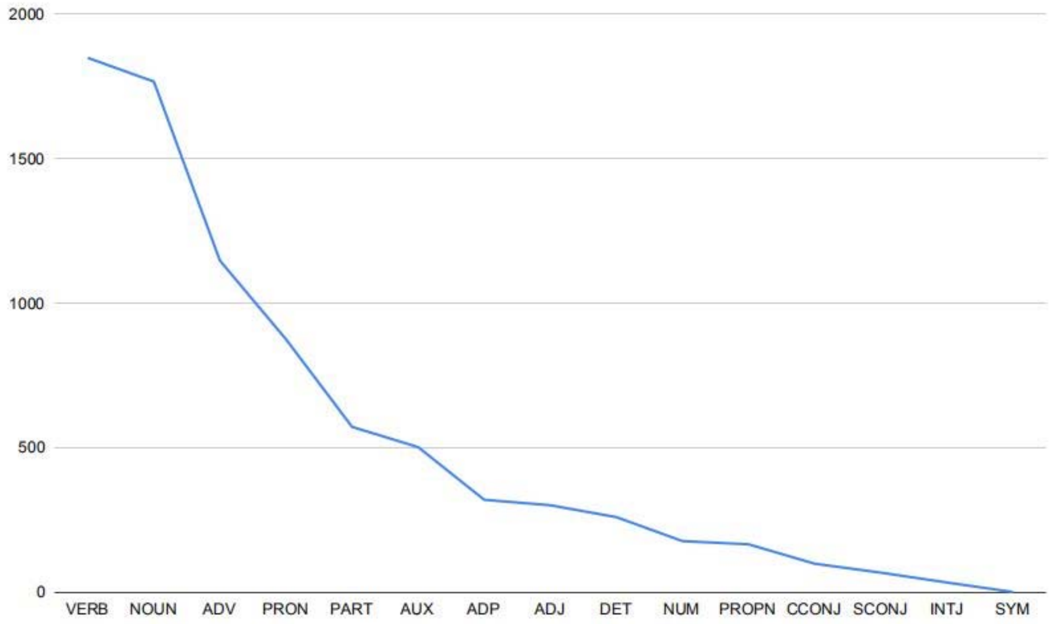
dislocated:vo (Dislocated object of verb-object compound)	--	FOB ((Fronting-object))?
mark:adv (Manner adverbializer 地 *de*)	dvpm (Manner 地 *de* modifier)	RAD ((Right adjunct))
mark:ext (得 *dé* in extent construction)	--	CMP ((Complement)) or RAD ((Right adjunct)) ??
mark:rel (Adjectival/complementizer /nominalizer 的 *de*)	cpm (< Complementizer)	RAD ((Right adjunct))
nsubj:pass (Nominal passive subject)	nsubjpass (Nominal passive subject)	FOB ((Fronting object))
obl:agent (Agent in passive phrase)	pass ↓ ((< Passive marker))	ADV ↓ ((Adverbial))
obl:patient (Object in 把 *bǎ* construction)	pobj ?	POB ↺ ((Preposition-object))
obl:tmod (Temporal nominal modifier)	tmod (Temporal modifier)	ADV ((Adverbial))

(( )) = Double parentheses indicate the relation covers other use cases not found in the Chinese UD label of the same row
< = The 'less than' symbol indicates the relation covers only a subset of the use cases in the Chinese UD label of the same row
( ) = Text in both single and double parentheses indicate the definition of the label
↺ = The dependency is in the opposite direction ↑ = the head is different ↓ = the dependent is different
? = Possible but uncertain due to insufficient information -- = unknown or no match found

Appendix 4. Distribution of the relations in the UD Chinese HK treebank



Appendix 5. Distribution of the parts of speech in the UD Chinese HK treebank


VERB
NOUN
ADV
PRON
PART
AUX
ADP
ADJ
DET
NUM
PROPN
CCONJ
SCONJ
INTJ
SYM