Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank

Tak-sum Wong
City University of Hong Kong
tswong-c@my.cityu.edu.hk

Kim Gerdes
Sorbonne Nouvelle, LPP (CNRS)
Paris, France
kim@gerdes.fr

Herman Leung
City University of Hong Kong
leung.hm@gmail.com

John Lee
City University of Hong Kong
jsylee@cityu.edu.hk

Abstract

This paper describes a new Cantonese-Mandarin parallel dependency treebank. We discuss the extent to which the treebank allows for comparative measures with the goal of quantifying structural differences between the two languages. After presenting syntactic differences between the two languages, we computed various frequency measures on the treebank. We present the results and discuss whether they reflect differences in text genre, differences in annotation scheme design, or actual structural differences. Finally, we compare the structural differences to previous accounts of the observed construction.

1 Introduction

Cantonese is part of the Yue dialect group which is spoken by more than 55 million people mostly in Canton, Hong Kong, Macao, the rest of the Pearl River Delta, and overseas Chinese communities. It is the "most widely known and influential variety of Chinese other than Mandarin" (Matthews & Yip 1994), and the early contact of Cantonese speakers with European explorers has given rise to the Western "Cantonese" pronunciations of some Chinese cities (e.g. *Canton*). Cantonese is not only used orally or in informal conversation, but also in the legislative councils in Hong Kong and Macao.

The special status of Hong Kong and Macao and the economic and educational importance of the region has made Cantonese a relatively well-studied and well-resourced language. A number of Cantonese corpora have already been tagged with part-of-speech (POS), including the Early Cantonese Tagged Database (Yiu 2012), the Hong Kong Cantonese Child Language Corpus (CANCORP, Lee et al. 1996), the Hong Kong Bilingual Child Language Corpus (Yip and Matthews 2007), the Hong Kong Cantonese Corpus (HKCanCor, Luke & Wong 2015), the Cantonese Chinese Corpus of Oral Narratives (CANON, Law et al. 2012), and the Hong Kong Mid-1990s Newspaper Column Corpus (Li et al. 2016). However, to our best knowledge, no syntactic treebank has been published prior to our work, neither phrase structure nor dependency based.

This paper presents the first parallel dependency treebank for Cantonese and Mandarin and analyzes statistical differences between the treebanks. The rest of the paper is organized as follows. The next section summarizes syntactic differences between Cantonese and Mandarin. Section 3 discusses the construction process of the treebanks. Section 4 presents statistical analyses on the treebank. Finally, Section 5 concludes.

2 Linguistic background

Cantonese and Mandarin are similar languages in most major respects, leaving aside pronunciation and grammatical particles. Some significant linguistic differences between the two languages are well-established (Ouyang 1993), including phonology, vocabulary, and in particular the rich Cantonese system of utterance particles. Some differences of grammatical structure have been described as well but, due to the absence of a Cantonese treebank and, even less so, of a parallel treebank, descriptions of structural differences could not be put on empirical grounds so far. We will show that some of these differences reflect measures that we can take on our treebank; for other phenomena our treebank does not yet provide enough data to assess significant differences.

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 266-275,
Pisa, Italy, September 18-20 2017

2.1 Double objects

Among the commonly known syntactic differences we have to cite is the canonical word order of monotransitive and ditransitive verb constructions, which is reversed compared to Mandarin: For a ditransitive verb, in Cantonese we have the following word order:
verb + direct object + indirect object.

畀	一枝花	我
Péi	*yātjīfā*	*ngóh*
give	a flower	1SG

‘Give me a flower.’

In Mandarin it is
verb + indirect object + direct object.

給	我	一枝花ㄦ
Gěi	*wǒ*	*yīzhīhuār*
give	1SG	a flower

‘Give me a flower.’

These two alternative constructions recall the English dative shift alternation.

2.2 Use of the object marker

For monotransitive verbs, the object marker (OM) being more prominent in Mandarin, the SOV order is more frequent in Cantonese. The same word order exists in Cantonese but is marked. It is used when the speaker wants to put stress on the object. The two competing Cantonese constructions are:

閂	咗	度	**門**	啦！
Sāan	*jó*	*douh*	***mùhn***	*lā!*
close	PERF	CLF	**door**	SFP

‘Close the door!’
PERF=perfective particle
CLF=classifier
SFP=sentence final particle

vs.

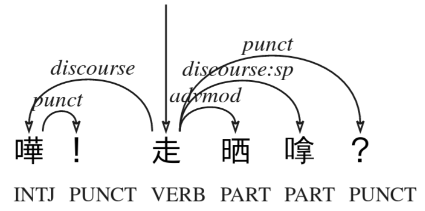
將	度	**門**	閂	咗	（佢）	啦！
Jēung	*douh*	***mùhn***	*sāan*	*jó*	*(kéuih)*	*lā!*
OM	CLF	**door**	close	PERF	(3SG)	SFP

‘the Door, close (it)!’

2.3 Post-verbal modifiers

Another notable difference of the two languages is the structure of post-verbal modifiers: Compare the following Cantonese sentence (Nr 0_189 of the parallel treebank) with its Mandarin counterpart.

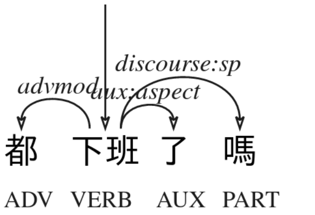
Cantonese:



Wa!	*Jáu*	***saai***	*làh?*
Wow	go	**all**	SFP

‘Wow! All of them have gone already’ / ‘They have all gone?’ / ‘They have all been released from duty?’

Mandarin:



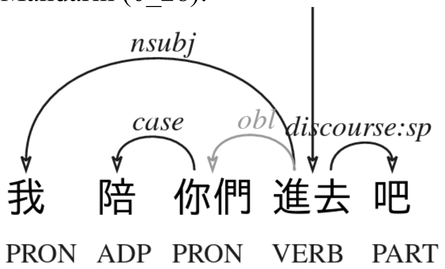
Dōu	*xiàbān*	*le*	*ma*
all	off-duty	ASP	SFP

The Cantonese post-verbal modifier 晒 *saai* ‘all’ is often considered as a quantifying verb-compound with the verb grammaticalizing to a quantifying particle that can translate as “additionally, also”. The Mandarin counterpart is an adverb in the standard preverbal position.

2.4 Coverb constructions

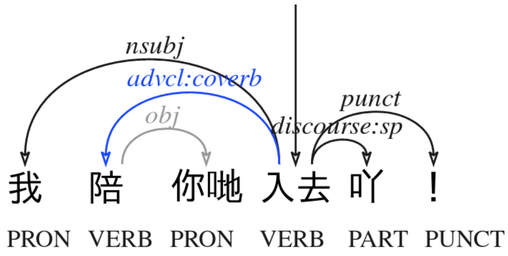
As pointed out by Francis and Matthews (2006), Cantonese coverbs are actually verbs, *e.g.* they can be used with aspect markers and verbal particles, In contrast, the Mandarin counterpart is rather a preposition – a preposition of verbal origin that has lost all of its verbal properties, except that it can still take a (prepositional) object.

Mandarin (0_28):



Wǒ	***péi***	*nǐmen*	*jìnqù*	*ba*
1SG	**accompany/with**	2PL	go.inside	SFP

Cantonese:



Ngóh ***pùih*** *léihdeih jahpheui ā*
1SG **accompany** 2PL go.inside SFP
‘Let me enter / go into the shop with you!’

Similarly, in beneficial constructions, the English preposition *for* is translated by the polysemous character for *give*. Its usage in Mandarin is quite grammaticalized and it is usually considered a preposition, Cantonese remaining more analytical. In our Mandarin UD guidelines, we introduce a specific sub-relation of advcl, *advcl:coverb*, to account for these constructions.

2.5 Expletives

A last well-known difference between Cantonese and Mandarin is the existence of expletives in Cantonese (annotated with the relation name *expl*), which are completely absent from Mandarin. An example is 佢 *kéuih* ‘3SG’. The pronoun is part of a grammatical construction and actually does not refer to anything or anyone, the condition for qualifying as expletive.[1]

大家 飲勝 **佢！**
Daaihgā jámsing ***kéuih***
everyone cheers **KEUHI**
‘Everyone! Cheers (to it)!’

我 不如 死 咗 **佢** 好過 啦!
Ngóh bātyùh séi jó ***kéuih*** *hóugwo lā*
1SG had.better die PERF **KEUIH** better SFP
‘It would be better for me to die.’

3 Treebank construction

Our corpus is based on television programs broadcast in Hong Kong (Lee, 2011). The Cantonese text is thus semi-planned spoken text. Cantonese TV dramas are widely distributed in southern China and beyond and mostly have Mandarin subtitles. The annotation is still ongoing and the texts that still await annotation are taken from movies that are distributed on Youtube, which will ultimately allow transforming this part of the treebank into a completely free language resource since the creators agreed to the distribution of the language data. The spoken Cantonese was transcribed with traditional Chinese characters by a native speaker of the language.

Although the subtitles were in traditional Chinese, we added a transcription in simplified Chinese as a separate feature. The reason being that we need both character sets: The simplified characters are necessary in order to apply parsing and segmentation tools. And we kept the traditional characters because the ongoing alignment is more straightforward with identical character sets and also because the Hong Kong residents who are working on the project are more used to traditional characters. Moreover, the projection from traditional to simplified characters is mostly one-to-one but for some characters many to one, and thus easier in the direction *traditional → simplified*.

The Cantonese transcription was done independently of the Mandarin subtitles. This has important consequences on the measures that we are able to take, because, as we will see, the treebank is not as strictly parallel as we had hoped because the subtitles are condensed and simplified versions of the Cantonese original.
The currently annotated part of the corpus consists of 569 parallel sentences. The treebank is sentence-aligned. As shown in Table 1, the spoken Cantonese sentences are longer than their counterpart of Mandarin subtitles.

Language	Number of tokens	Average sentence length
Mandarin	4149	7.29
Cantonese	5428	9.54

Table 1: Corpus data

3.1 The UD annotation scheme

For the annotation of the parallel treebank, we decided to follow the Universal Dependency (UD, de Marneffe et al., 2014; Nivre et al., 2016) annotation scheme, as this allows the comparison of our resource also with external treebanks. However, even for Mandarin, no annotation guide existed, and the first UD v1 Mandarin treebank does not come with any explanation of the annotation choices and its annotation is, unsurprisingly, quite heterogeneous.

The Mandarin UD v1 annotation guide was explicitly developed for the UD dependency annotation of the Mandarin side of our corpus. Leung et al. (2016) describe the underlying discussions and choices, in particular for Chinese idiosyncrasies like classifiers, aspectual and sentence final particles, and light verb as well as serial

[1] For further details and examples see http://universaldependencies.org/yue/dep/expl.html

verb constructions. In accordance with discussions around the development of this Mandarin annotation guide, UD v2 explicitly takes into account a specific *clf* 'classifier' relation, which is a unique type of syntactic relations that only exists for languages that have classifiers – Mandarin being the first language with this feature that is described in UD.

The UD v1 guide has been completed during the ongoing annotation experience and then adapted to v2. The Mandarin-specific part of the UD documentation is currently one of the most complete language specific annotation descriptions[2]

The similarity of Cantonese and Mandarin makes it reasonable to conceive the Cantonese annotation guide on the basis of the Mandarin guide, with modifications wherever necessary. The development of this guide is work in progress.

The whole semi-automatic annotation process is done in the Arborator annotation tool (Gerdes 2013), which allows blind and open annotation by multiple users as well as integrated parser bootstrapping possibilities

3.2 Outline

UD has been conceived with a double objective: The parallel construction of the treebanks facilitates the developments of parsers and other NLP tools. And, more importantly for the present study, it allows studies in empirical comparative syntax. There are some caveats to this claim, some of which we will discuss later. But any comparative measure on the current UD treebanks will always measure either structural differences, genre differences of the underlying corpus, differences in the design of the annotation scheme, or annotation errors and incoherences of course. Our corpus is, at least partly different in this aspect: Being a parallel treebank, the content of both treebanks is identical and any ascertained difference should be attributed to a structural difference. Alas, as we have mentioned before, this is not completely true, as the Mandarin subtitles are not precise translations of the original Cantonese words. Therefore, the measured differences can always either be an actual syntactic difference, or rather a difference of genre: The genre of spoken texts in TV dramas vs. the genre of subtitles in "Translationese" – although the pure informational content is mostly identical.

The measured differences between the two sides of the parallel treebank that cannot easily be attributed to the genre variation may either be new to us or corroborate known syntactic differences between the two languages.

4 Statistical measures

This section first presents the statistical measures that will be used to assure the validity of the significance of the observations (Section 4.1). Further, various difference measures based on the POS distribution will be presented and discussed (Section 4.2). Then we move on to differences in the functional distribution (Section 4.3) and finally we mix categorical and functional information (Section 4.4). After a short presentation of dependency directional measures (Section 4.5), we will conclude with an outlook on the ongoing annotation and alignment process.

4.1 Fisher Test and Specificity

In order to distinguish significant from insignificant over- and under-representation of features of our parallel treebank, we systematically apply the exact Fisher test which is based on the cumulative hypergeometric distribution. The null hypothesis is that the size of the two corpora as well as the number of total words having a specific category (or syntactic function) being fixed, the actually observed number of occurrences is due to chance. The p-value measures the probability that the observed frequency (or more occurrences if the number is already over-represented or less if already under-represented) actually occurred. To make the probabilities more readable, we transform them in *specificity* values (Lebart et al. 1991): specificity=$-\log_{10}(p)$ if the observed frequency is higher than the expected value and $\log_{10}(1-p)$ if the frequency is lower than expected. The expected value is the equidistribution of categories and functions into the two corpora depending on the size of the corpora and the frequency of the categories and functions. This is a well-established method in textual statistics, but still quite rarely used in syntactic comparisons.

4.2 Categorical differences

Concerning the POS distribution we observe the following variation between Cantonese and Mandarin. The first line of Table 2 can be read as follows: Cantonese contains 999 of the total 1344 PUNCT(uation) tokens in our two treebanks. The positive Specificity value indicates

[2] The annotation guidelines that we have developed can be accessed at: http://universaldependencies.org/zh/overview/introduction.html

that PUNCT is over-represented in Cantonese. The probability that this is due to chance is very low: $1/10^{31}$.

Type	Specificity	Cantonese	Total
PUNCT	31	999	1344
INTJ	23	97	97
PART	10	619	898
X	5	20	20
AUX	0	246	428
CCONJ	0	18	33
SCONJ	0	23	41
ADJ	-1	97	186
NOUN	-1	801	1449
NUM	-1	54	104
PROPN	-1	84	155
DET	-4	60	144
VERB	-4	347	688
PRON	-5	462	915
ADP	-8	93	239
ADV	-11	511	1080

Table 2: POS frequencies by specificity

Inversely, the last row of Table 2 indicates the following observation: Cantonese has only 511 of the total of 1080 ADV(erbial) tokens. This is less than statistically expected if the POS were distributed evenly, given that the Cantonese part of the corpus is larger. The probability that the observed frequency difference is due to chance is $1/10^{11}$. The upper shaded (green) rows of the table thus show significant over-representation of categories in Cantonese, the lower shaded (red) rows show significant under-representation. The unshaded rows have over- or under-representation of order 0 or 1 (p~1/10) and thus non-significant differences. The significantly lower frequency of adverbs in Cantonese is likely due to the prominence of Cantonese post-verbal particles where in Mandarin adverbs are often used to express the same meaning. For instance, for the progressive aspect, in Mandarin the adverb *zhèngzài* 正在 is used (*zhèngzài* + *V*) where the Cantonese counterpart is V-*gán* in which *gán* 緊 is a post-verbal aspect particle. (Also cf. section 2.3)

We see that the Cantonese treebank was not only punctuated very differently than the Mandarin subtitles. The Cantonese side contains all the observed interjections of the whole parallel treebank as well as a much higher frequency of particles. This underlines again that the subtitle translation is actually a condensed, not to say impoverished, version that lacks many of the oral features of the spoken original. The fact that the POS tag X (words where annotators cannot determine a POS, like the prefix *a* 阿) only appear in Cantonese can be attributed to possible disagreements between the annotators which may be due to the oral character of the transcription as well as to the underdeveloped formal grammars of Cantonese – making the annotation task harder.

Further, we observe the expected under-representation of ADP(ositions) in Cantonese due to the verbal character of many Cantonese equivalents of Mandarin prepositions, as discussed in section 2.4. It remains to explain why verbs are nonetheless also under-represented in Cantonese.

The under-representation of PRON(ouns) in Cantonese is unexpected. This may be an actual linguistic difference between the two languages or it may be due to the less oral character of the Mandarin translation compared to the Cantonese transcriptions, leading to less pronoun dropping. This will have to be examined further.

4.3 Functional differences

Table 3 shows the significant differences in the distribution of syntactic functions, partly corresponds to what has been observed for the POS (e.g. the high frequency of *punct*, *discourse*, and *discourse:sp* = "sentence final particle" relations), but also shows a few more interesting variations: The current Mandarin annotation does not contain any *advcl:coverb* relations, which is due to differences in annotation, but which nonetheless reveals a significant structural difference between the languages: The Mandarin prepositions are of verbal origin but have lost all verbal properties whereas their Cantonese counterparts can still be modified by verbal articles and have thus to be tagged and annotated differently (see section 2.4). The UD annotation scheme handles prepositions as case-markers, and thus as dependent from their argument, i.e. what is commonly called a prepositional object. This results in UD's

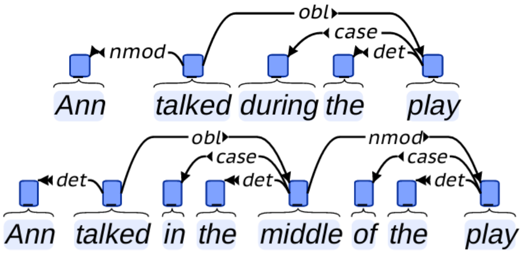
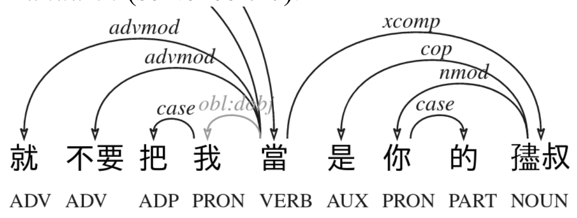


Figure 1: Analyses of two (semantically full) prepositions in UD 2.0 English, the first being a simple and the second a complex preposition

infamous “Turkish” analysis of English prepositions (Chris Manning, 2016, personal communication). Figure 1 shows the situation for English (example taken from Gerdes & Kahane 2016, updated to UD 2.0).

The following pair of sentence segments illustrates this point for Chinese. The 1st person singular pronoun in the Mandarin tree 我 *‘wǒ’* is an obl:dobj that has a case-marker. In the Cantonese equivalent, what has been analyzed as a (verbal) preposition in Mandarin is now a coverb, which takes its argument as a regular direct object.

Mandarin (sentence 0-7)*:*

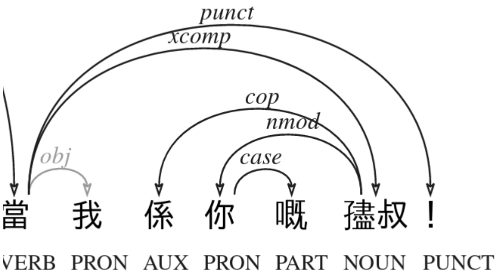


Jiù	*bùyào*	*bǎ*	*wǒ*	*dàng*	*shì*
then	do.not	OM	1SG	treat	COP

nǐ	*de*	*nāishū*
2SG	REL	youngest.uncle.on.paternal.side

‘Don’t treat me as your uncle.’

Cantonese:



Dong	*ngóh*	*haih*
treat	1SG	COP

léih	*ge*	*lāaisuk*
2SG	REL	youngest.uncle.on.paternal.side

‘Treat me as your uncle.’

We end up with structurally very different trees for a simple categorical choice. Note that the proximity between verbs and preposition is not reserved to Chinese. The English *during* or the French equivalent *pendant* are similar cases where the verbal character of the preposition is still visible.

Alternatively, we could have decided to treat all Cantonese coverbs as prepositions, so that the Cantonese trees would be in line with the Mandarin ones. This is a difficult choice as UD seeks “to maximize parallelism by allowing the same grammatical relation to be annotated in the same way across languages, while making enough crucial distinctions to differentiate constructions that are not the same.” (Nivre 2015 and UD home-

Type	Spec	Cantonese	Total
punct	31	1002	1345
discourse	26	204	226
discourse:sp	11	443	619
advcl:coverb	9	40	40
det	3	193	286
goeswith	2	25	33
advmod:df	1	12	17
aux:aspect	1	80	125
cop	1	76	125
appos	0	27	45
csubj	0	15	24
iobj	0	1	3
mark:dev	0	1	1
obl:agent	0	1	3
obl:clf	0	2	3
obl:poss	0	2	4
acl	-1	34	73
amod	-1	40	75
aux	-1	90	171
aux:pass	-1	0	2
case:loc	-1	26	52
cc	-1	17	33
clf	-1	47	88
mark	-1	38	76
nsubj:pass	-1	0	3
nummod	-1	53	99
obl:tmod	-1	83	154
parataxis	-1	84	161
vocative	-1	69	128
advcl	-2	91	184
nmod	-2	99	204
obj	-2	393	726
mark:rel	-3	20	56
nsubj	-3	362	707
xcomp	-3	64	140
dislocated	-4	62	148
obl	-5	58	147
ccomp	-6	56	145
advmod	-7	541	1087
obl:dobj	-7	0	18
case	-14	80	245

Table 3: complete dependency relation frequencies ordered by specificity

page. And although prepositions in English are considered by any syntactic analysis that we are aware of to be "crucially" different from case markers (Osborne 2015), UD decided to treat them just like Turkish case markers, leading to greater similarity between Turkish and English and at the same time to the structurally very different trees for simple and complex prepositions (Figure 1)

A good syntactic annotation scheme would allow for slight structural differences to be reflected by slight differences in the annotation, for example in the case of Cantonese coverbs by a different categorization of the coverb, once as a verb and once as a preposition, but with identical dependency structures in both treebanks. The "Turkish" analysis of prepositions, on the contrary, triggers a structural upheaval, for a small real difference: A "catastrophe" in a strictly mathematical sense of Thom's catastrophe theory (Saunders 1980, Gerdes & Kahane 2016), i.e. a brutal structural change in a continuum. This results in measures of important differences where there are few (between Mandarin and Cantonese for example), and in the absence of annotation differences where syntactic differences actually occur (e.g. English prepositions vs. Turkish case markers).

The UD annotation scheme obliges all dependency relations to be taken from a fixed set of 37 functions but it allows for the creation of idiosyncratic sub-relations when needed by a given language. The sub-relations are separated by a colon from the main relation: *relation:subrelation*. When grouping together subrelations, we obtain Table 4, a simpler table with similar significant variations between Cantonese and Mandarin. Concerning the adverbial clause (*advcl*) relation, we see that its distribution is no longer significantly different between the two languages: Mandarin had more simple *advcl*, Cantonese more coverb constructions which adds up to an equal distribution.

Type	**Spec**	**Cantonese**	**Total**
punct	31	1002	1345
discourse	27	647	845
det	3	193	286
goeswith	2	25	33
cop	1	76	125
advcl	0	131	224
appos	0	27	45
aux	0	170	298
csubj	0	15	24
iobj	0	1	3
acl	-1	34	73
amod	-1	40	75
cc	-1	17	33
clf	-1	47	88
nummod	-1	53	99
parataxis	-1	84	161
vocative	-1	69	128
nmod	-2	99	204
obj	-2	393	726
mark	-3	59	133
xcomp	-3	64	140
dislocated	-4	62	148
nsubj	-4	362	710
advmod	-6	553	1104
ccomp	-6	56	145
obl	-6	146	329
case	-14	106	297

Table 4: simple dependency relation frequencies ordered by specificity (simple meaning that subrelations are grouped under the main relation)

4.4 Mixed measures

When grouping together the syntactic function and the POS of the dependent token, we obtain 128 classes of function-POS pairs. Although the small size of our current parallel corpus makes most differences fall under the significance threshold, some couples are significantly over- and under-represented. See Table 5 for details.

We observe for example that Cantonese particles are mostly in discourse or advmod relations whereas Mandarin particles are mark (~verbal complementizers) and case markers (~prepositions).

Since UD v2.0, the *dislocated* relation is used for objects in a non-canonical position "that do not fulfill the usual core grammatical relations of a sentence" (UD page for the *dislocated* relation[3]), so all the *obj* and *obl* relations in the above list are actually post-verbal. Since the Cantonese data is more oral, the over-representation of objects could also partially be due to this distinction and not to an actual difference in the valency structures of the observed verbal objects.

[3] It is not completely clear what is actually meant by "fulfilling the core grammatical relation" because a dislocated object usually fills the valency slot of the verbal governor. Mimicking what has been done for English and French, we decided to annotate preverbal objects with the *dislocated* relation.

Type	Spec	Cantonese	Total
punct→PUNCT	31	998	1341
discourse→INTJ	23	97	97
det→NOUN	19	126	135
discourse→PART	18	516	692
advmod→PART	10	44	44
det→PRON	2	7	7
goeswith→NOUN	2	15	18
vocative→X	2	7	7
…			
acl→VERB	-2	32	70
dislocated→NOUN	-2	43	92
nmod→PRON	-2	71	146
nsubj→NOUN	-2	87	178
obj→NOUN	-2	266	505
obl→PROPN	-2	2	10
xcomp→VERB	-2	49	110
mark→PART	-3	25	68
nsubj→PRON	-3	252	490
obl→NOUN	-3	120	247
det→DET	-4	60	144
case→PART	-5	30	89
ccomp→VERB	-5	44	119
dislocated→ADV	-5	0	13
obl→PRON	-6	18	63
advmod→ADV	-10	472	1004
case→ADP	-10	73	204

Table 5: selection of dependency-POS couples, ordered by specificity

If we go one step further, we can measure triples *POS–func→ POS*. The two treebanks contain more than 300 of these triples, the two most frequent ones, with more than 700 occurrences being *VERB–punct→PUNCT* and *VERB–advmod→ADV*.

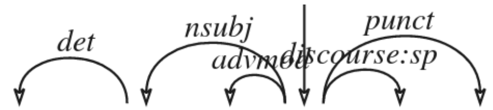
The most significantly over-represented Cantonese triples are shown in Table 6.

The significant over-representation of *NOUN–det→NOUN* relations in Cantonese may seem surprising and does not seem to follow directly from the POS distribution. Note first that the fixed UD POS tag-set does not include a specific category for classifiers which are therefore tagged as nouns. What we are actually observing here is that bare classifier noun phrases [CLF NOUN] is a common Cantonese strategy for definite NP constructions. In Cantonese only [CLF NOUN] and [DET CLF NOUN] are possible for definite NPs. In Mandarin we have [NOUN], [DET NOUN], or [DET CLF NOUN].[4]

Type	Spec	Cantonese	Total
VERB-punct→PUNCT	24	595	781
INTJ-punct→PUNCT	22	93	93
NOUN-det→NOUN	19	126	135
VERB-discourse→INTJ	15	64	64
VERB-discourse→PART	12	369	503

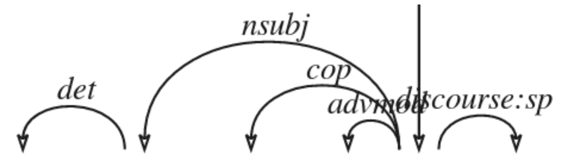
Table 6: The most over-represented triples POS – dependency – POS on the Cantonese side of the parallel treebank, ordered by specificity

Cantonese (sentence 0_2):



個	看更	新	唻	㗎	。
NOUN	NOUN	ADJ	VERB	PART	PUNCT
Go	*hōn'gāang*	*sān*	*làih*	*ga*	
CLF	watchman	new	arrive	SFP	

Mandarin:



那	保安員	是	新	來	的
DET	NOUN	VERB	ADJ	VERB	PART
Nà	*bǎoānyuán*	*shì*	*xīn*	*lái*	*de*
CLF	watchman	COP	new	arrive	SFP

On the lower edge of the table, the most typically Mandarin triples are these:

VERB-advmod→ADV	-10	332	729
AUX-ccomp→VERB	-14	0	38

Table 7: The most significantly over-represented triples POS – dependency – POS on the Mandarin side of the parallel treebank

In common copula constructions, UD imposes the analysis of the copula verb as the de-

[4] Note that [CLF NOUN] is also possible in Mandarin, but only in post-verbal position, and it can only have an indefinite interpretation, hence it occurs much less frequently than in Cantonese. In Cantonese, [CLF NOUN] can occur in both preverbal and postverbal position, but in preverbal position it must be definite; in postverbal position, it can be ambiguous between definite and indefinite.

pendent of the semantically full element, which is commonly a noun or an adjective. In the new UD v2 annotation scheme however, the auxiliary is considered the head of the construction if the semantically full argument is a verb itself, the copula verb becomes the head of the construction, a decision which attempts to avoid cases of embedded multiple auxiliary constructions where the subject can no longer be unequivocally attributed to its governor. This explains the existence of the *AUX–ccomp→VERB* triple, but it does not explain why this construction is over-represented in Mandarin. This will have to be explained by returning on the actual parallel data where the *AUX–ccomp→VERB* triple must have a structurally different translation in Cantonese.

4.5 Directional measures

A final set of measures on the treebank is based on the direction of the dependency link:

name	***advmod***	***aux***	***obj***	***obl***
Cantonese	13,74	48,82	100	28,08
Mandarin	3,81	35,16	100	19,67

Table 8: Percentage of right-pointing relations by syntactic function: A selection of functions

This kind of measures has been used in various treebank analysis methods, in particular in typological research, where the direction of the head-daughter relations has been shown to correlate with many important language features (Liu 2010, Chen & Gerdes 2017).

Here we just briefly want to point to a few aspects that have been mentioned above: We see that our annotation scheme only has objects to the right of its verbal governor – other positions would be annotated as *dislocated*. For the oblique verbal argument, however, we observe an important difference between Cantonese and Mandarin: Mandarin has around 20% of its oblique arguments to the right of their governor – Cantonese has 10% more, corresponding to the aforementioned structural preferences.

The higher number of right-branching *advmod* and *aux* relations in Cantonese, however, does not follow directly from the known language differences and should be explored further, preferably on more, and if possible, less genre dependent parallel data.

5 Conclusion

This article presents a method of empirical comparative syntax using statistical measures on a comparatively small sentence-aligned parallel dependency treebank. The specificity measurements, based on the exact Fisher test, are well-adapted to small corpora because the alternative test for categorical data, the approximating χ^2 test, gives incorrect results for very small (and very frequent) occurrences (compared to the size of the corpus) – and the frequencies of most words in a corpus are very low.

The significant observations can often be explained by actual differences in the language structure or at least in the language annotation scheme. Since the corpus is parallel, the differences are not due to different vocabulary etc., but the subtle genre differences on the two sides of our treebank (transcription vs subtitle) remain very visible in the resulting measures.

We can see that Cantonese has significant structural differences with its Mandarin counterpart, although some of these differences are reinforced by the UD annotation scheme while other actual structural differences may have remained hidden from our statistical analysis. Inversely, however, not all well-known structural differences between the languages can be put under scrutiny by means of the parallel treebank. The expletive, for example, is absent from our corpus – pointing to the fact that frequently discussed phenomena are not necessarily frequent syntactic phenomena. The specificity measure allows ordering the observed differences by statistical importance, the degree of astonishment, thus empirically guiding the research to actual hotspots of syntactic variation.

The annotation choices we face with different stages of preopositional grammaticalization in a parallel or comparable treebanks can be seen as part of a more general question about the goal of the syntactic annotation: The UD choice to favor similar structures whenever possible leads to skewed typological similarity measures. Future UD schemes should be evaluated as to the extent that they allow avoiding catastrophes and capturing similarities between closely related structures.

The ongoing word alignment of the parallel treebank will soon allow for more precise queries concerning the differences or similarity between the two languages. But just like for the annotation, the word alignment, too, is already a structural choice (one-to-many alignments?, one-to-zero alignments?) that determines which results can finally be extracted. Ideally the word-alignment would allow for complementary measurements that cannot be obtained on the sole sentence aligned parallel treebank. Work in progress on a parallel treebank online query tool could also benefit from the integration of these types of statistical measures. It would allow to not only search for and count pre-discovered structural discrepancy, but rather permit exploring interesting facts hidden in the raw data.

Acknowledgments

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 N° 35322RG); and by two Strategic Research Grants (Project No. 7004494 and No. 7004736) from City University of Hong Kong.

References

Chen, Xinying, and Kim Gerdes. "Classifying Languages by Dependency Structure: Typologies of Delexicalized Universal Dependency Treebanks", *Depling*, 2017

David C. S. LI, Cathy S. P. WONG, Wai Mun LEUNG and Sam T. S. WONG. "Facilitation of Transference: The Case of Monosyllabic Salience in Hong Kong Cantonese" Linguistics, Vol. 54(1), pp. 1−58, January 2016.

Francis, Elaine J., and Stephen Matthews. "Categoriality and object extraction in Cantonese serial verb constructions." *Natural Language & Linguistic Theory* 24.3 (2006): 751-801.

Gerdes, Kim. "Collaborative Dependency Annotation." *Depling*, 2013.

Gerdes, Kim, and Sylvain Kahane. "Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies." *LAW X (2016) The 10th Linguistic Annotation Workshop*: 131. 2016.

Law SP, Kong APH, Lee A, Lai CT, Lam VVV. 2012. "Cantonese Chinese corpus of oral narratives (CANON) with morphological tagging: a preliminary report." Presented in the *Workshop on Innovations in Cantonese Linguistics (WICL)*, Columbus, OH., 16-17 March 2012.

Lebart, Ludovic, André Salem, and Lisette Berry. "Recent developments in the statistical processing of textual data." *Applied Stochastic Models and Data Analysis* 7.1 (1991): 47-62.

Leung, Herman, Rafaël Poiret, Tak sum Wong, Xinying Chen, Kim Gerdes, and John Lee "Developing Universal Dependencies for Mandarin Chinese." *The 12th Workshop on Asian Language Resources*. 2016.

Lee, John. Toward a Parallel Corpus of Spoken Cantonese and Written Chinese. In *Proc. 5th International Joint Conference on Natural Language Processing* (IJCNLP), 2011.

Lee, Thomas H. T. and Colleen Wong. 1998. CANCORP: the Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* vol. 27, no. 2, pp. 211-228.

Liu, Haitao. "Dependency direction as a means of word-order typology: A method based on dependency treebanks." *Lingua*, 120.6 (2010): 1567-1578.

Luke, Kang-Kwong, & Wong, May L-Y. 2015. The Hong Kong Cantonese Corpus: design and uses. *Journal of Chinese Linguistics* 25 (2015): 309-330

Matthews, Stephen and Virginia Yip. (2011) *Cantonese: A comprehensive grammar*. New York: Routledge.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*: 4584-4592.

Nivre, Joakim. "Towards a Universal Grammar for Natural Language Processing." *CICLing (1)* 2015 (2015): 3-16.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016a. Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 1659-1666.

Osborne, Timothy. "Diagnostics for Constituents: Dependency, Constituency, and the Status of Function Words." *Depling*, 2015.

Ōuyáng, Juéyà. (1993) 普通話廣州話的比較與學習 *Pŭtōnghuà Guăngzhōuhuà de bĭjiào yŭ xuéxí* (The comparison and learning of Mandarin and Cantonese). Peking: China Social Science Press.

Saunders, Peter T. *An introduction to catastrophe theory*. Cambridge University Press, 1980.

Yip, Virginia and Stephen Matthews. (2000) Syntactic transfer in a bilingual child. Bilingualism: Language and Cognition 3.3, 193-208

Yiu Yuk Man. Early Cantonese Tagged Database, presented at the *Workshop on Early Cantonese Grammar*, Dec 14 2014, Hong Kong: HKUST.