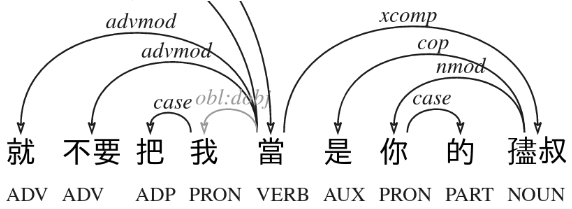
infamous “Turkish” analysis of English prepositions (Chris Manning, 2016, personal communication). Figure 1 shows the situation for English (example taken from Gerdes & Kahane 2016, updated to UD 2.0).

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.

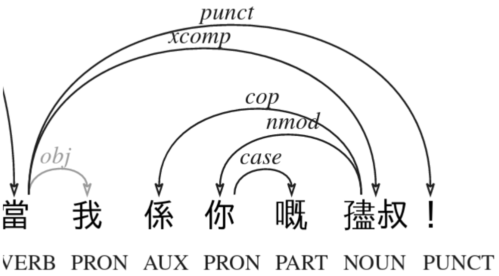
Mandarin (sentence 0-7)*:*


advmod
advmod
case
obl:dobj
xcomp
cop
nmod
case
就 不要 把 我 當 是 你 的 孻叔
ADV ADV ADP PRON VERB AUX PRON PART NOUN

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	Can-tonese	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*.

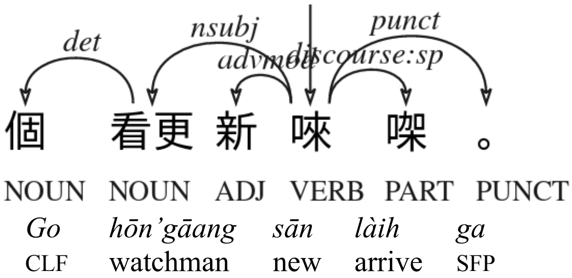
The most significantly over-represented Cantonese triples are shown in Table 6.

Type	Spec	Can-tonese	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

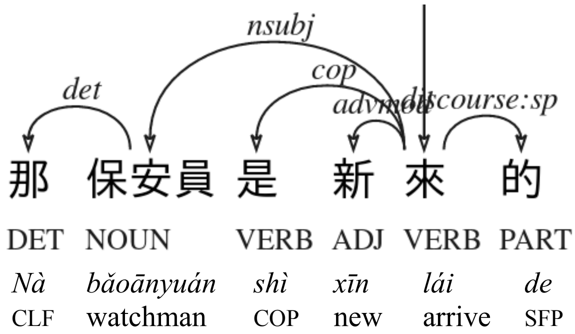
Table 6: The most over-represented triples POS – dependency – POS on the Cantonese side of the parallel treebank, ordered by specificity

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]

Cantonese (sentence 0_2):



Mandarin:



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 *Pǔtōnghuà Guǎngzhōuhuà de bǐjiào yǔ 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.