Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank

Tak-sum Wong*, Kim Gerdes[+], Herman Leung*, John Lee*

*Department of Linguistics and Translation
City University of Hong Kong

[+]Sorbonne Nouvelle, LPP (CNRS)
Paris, France

Introduction

- Cantonese, a Sinitic language, spoken by 55M people mostly in Canton, Hong Kong, Macao.
  “Cantonese is the most widely known and influential variety of Chinese other than Mandarin” (Matthews & Yip 1994)
- 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 POS-tagged corpora exist but no syntactic treebank has been published.
- We are presenting the first parallel dependency treebank for Cantonese and Mandarin and analyze the statistical differences.

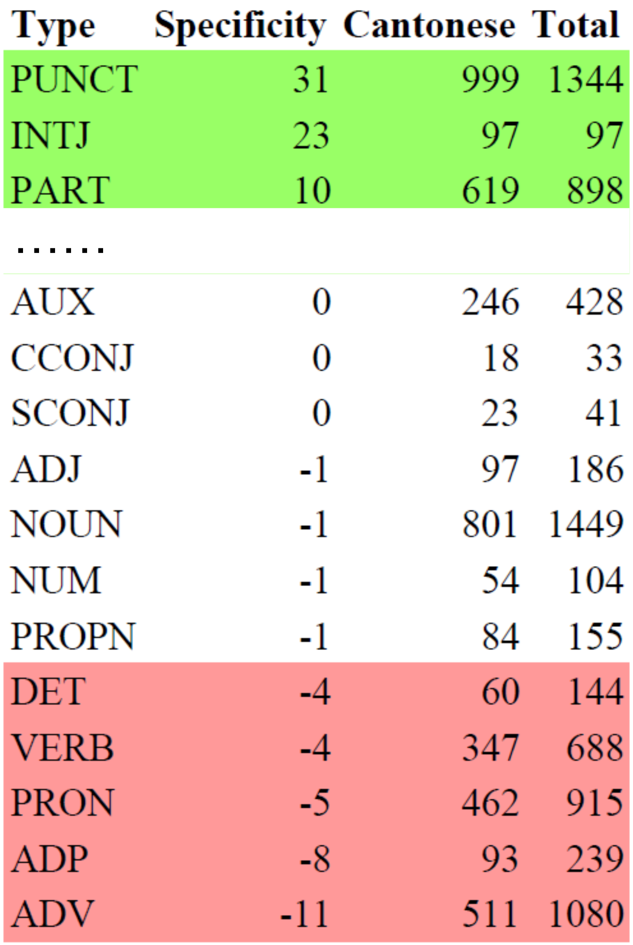
Treebank Construction

- Annotation scheme was adapted from existing UD guidelines for standard Chinese (Leung et al., 2016)
- Source Material: Hong Kong television programmes, with Mandarin subtitles
- Size: 569 parallel sentences
- Sentence-aligned
- Semi-planned spoken text
- Cantonese transcription was done independently of Mandarin subtitles
- Subtitles are always condensed, and simplified dialogues
- Treebank is not as strictly parallel

Language	#tokens	avg sent length
Mandarin	4149	7.29
Cantonese	5428	9.54

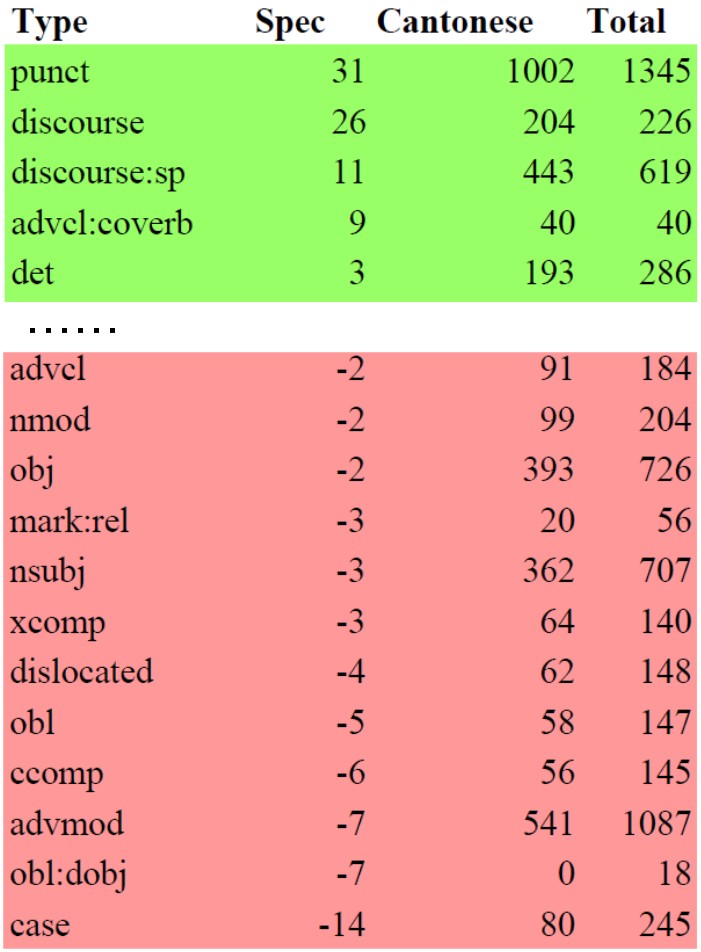
Statistical Measures

Categorical differences



Type	Specificity	Cantonese	Total
PUNCT	31	999	1344
INTJ	23	97	97
PART	10	619	898
……			
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

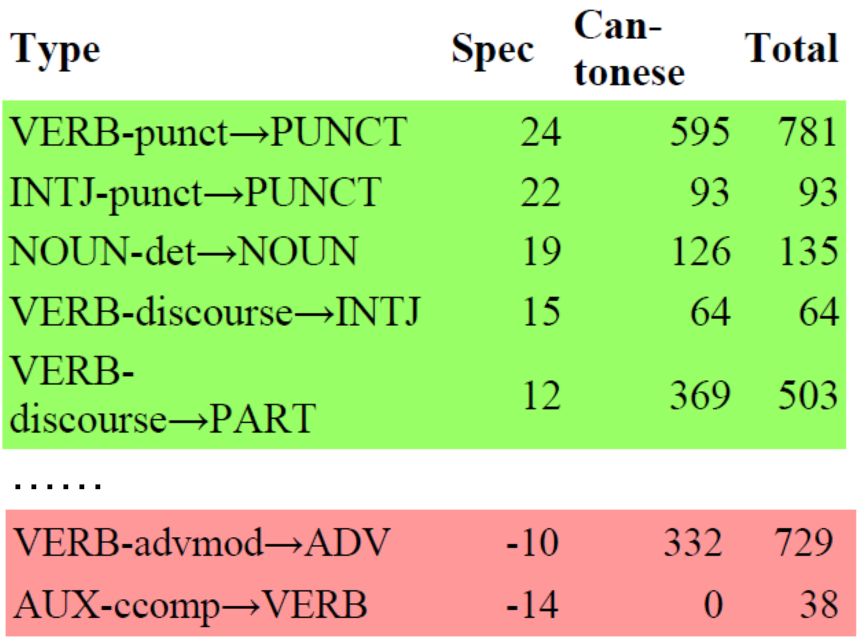
Functional measures



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
……			
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

Statistical Measures

Mixed measures



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

Directional measures

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

Artefacts vs. typology

- Parallel corpus, but:
  - Artefacts :
    - Different conventions

      → ***punct*** much more frequent in Cantonese
    - Translationese (genre)

      → ***INTJ*** much more frequent in Cantonese
  - Typology :
    - All points without explanation as artefact
      - Some conscious annotation choices
      - Some discoveries post-annotation

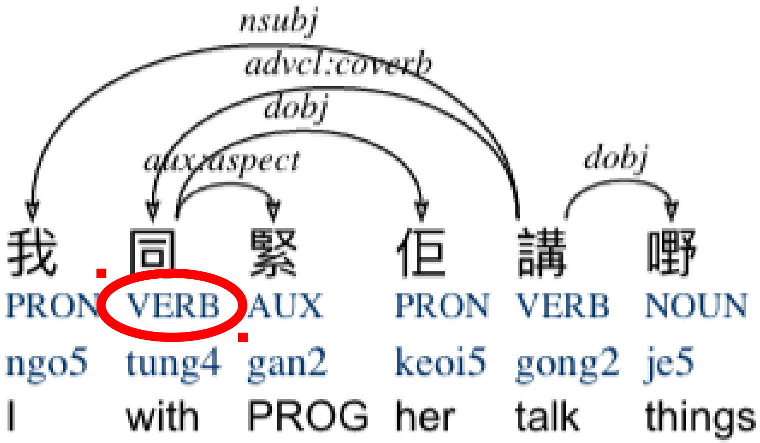
Preposition and (co)verb

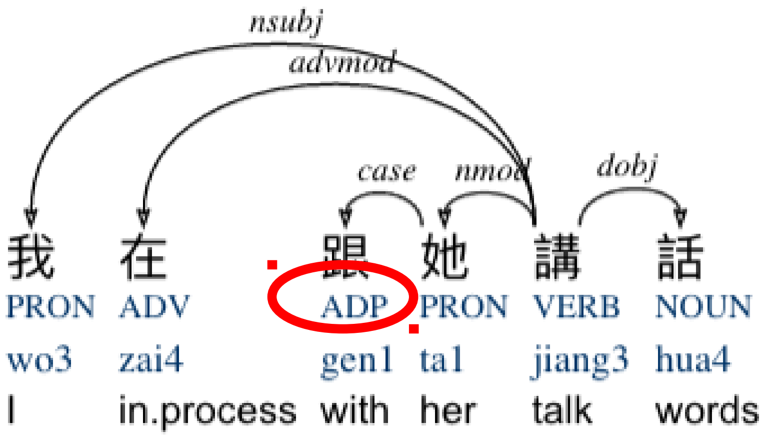
ADP	-8	93	239

advcl:coverb	9	40	40

case	-14	80	245

- Cantonese coverb is tagged as VERB+advcl:coverb
- Mandarin coverb is tagged as ADP (preposition) +case


nsubj
advcl:coverb
dobj
aux:aspect
dobj
我 同 緊 佢 講 嘢
PRON VERB AUX PRON VERB NOUN
ngo5 tung4 gan2 keoi5 gong2 je5
I with PROG her talk things


nsubj
advmod
case
nmod
dobj
我 在 跟 她 講 話
PRON ADV ADP PRON VERB NOUN
wo3 zai4 gen1 ta1 jiang3 hua4
I in.process with her talk words

Cantonese

‘I am talking with her’

Mandarin

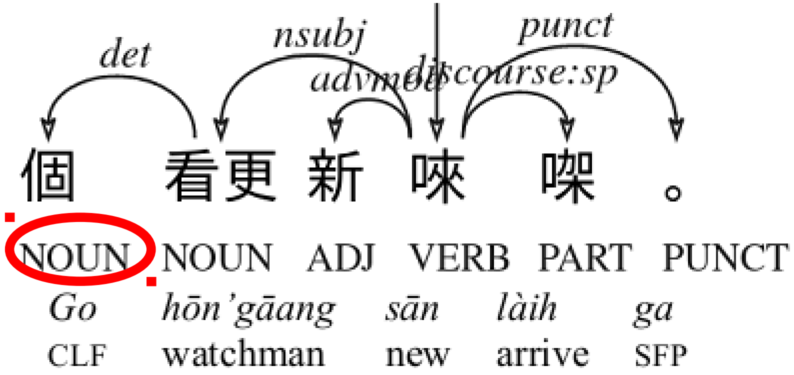
Noun(classifier) and determiner

- – “Bare classifier” construction in Cantonese: [classifier + noun] as definite NP
- – Aligned to a Mandarin demonstrative

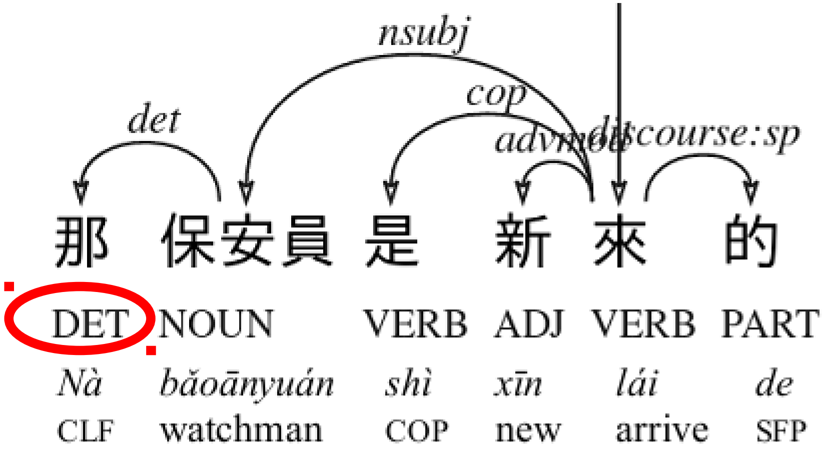
NOUN-det→NOUN	19	126	135

DET	-4	60	144

Cantonese (sentence 0_2):

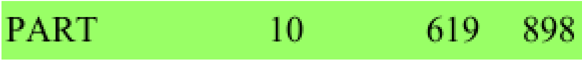


Mandarin:

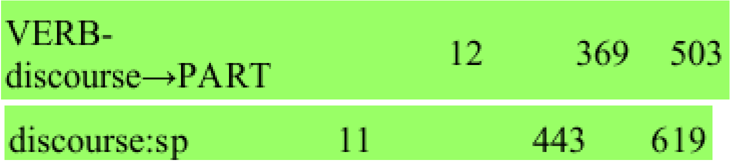

nsubj
det
cop
advmod
discourse:sp
那 保安員 是 新 來 的
DET NOUN VERB ADJ VERB PART
Nà bǎoānyuán shì xīn lái de
CLF watchman COP new arrive SFP

Sentence particle and adverb

– Some Cantonese sentence particles correspond to Mandarin adverbs



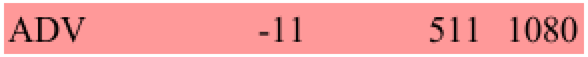
PART	10	619	898



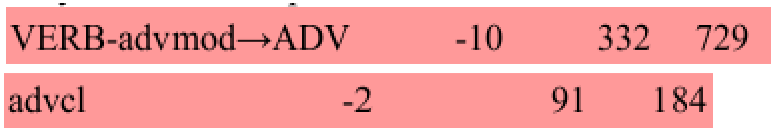
VERB-discourse→PART	12	369	503
discourse:sp	11	443	619

Cantonese	食	咗	凍	嘢	先 /PART
	eat	**PRF**	cold	thing	**first**

Mandarin	先 /ADV	吃	冷	的
	first	eat	cold	**NOM**



ADV	-11	511	1080



VERB-advmod→ADV	-10	332	729
advcl	-2	91	184

‘Eat the cold [things] first’

Conclusions

- A method of empirical comparative syntax using statistical measures on a sentence-aligned parallel dependency treebank.
- Significant observations can be explained by actual differences in the language structure.
- subtle genre differences on the two sides of our treebank: transcription vs subtitle is still visible

On-going Work

- Development of word alignment between Mandarin and Cantonese
- Transcribe materials distributed on Youtube for free language resource
- Analysing other constructions showing asymmetric difference between these two languages
- Application: for teaching Cantonese as a foreign language

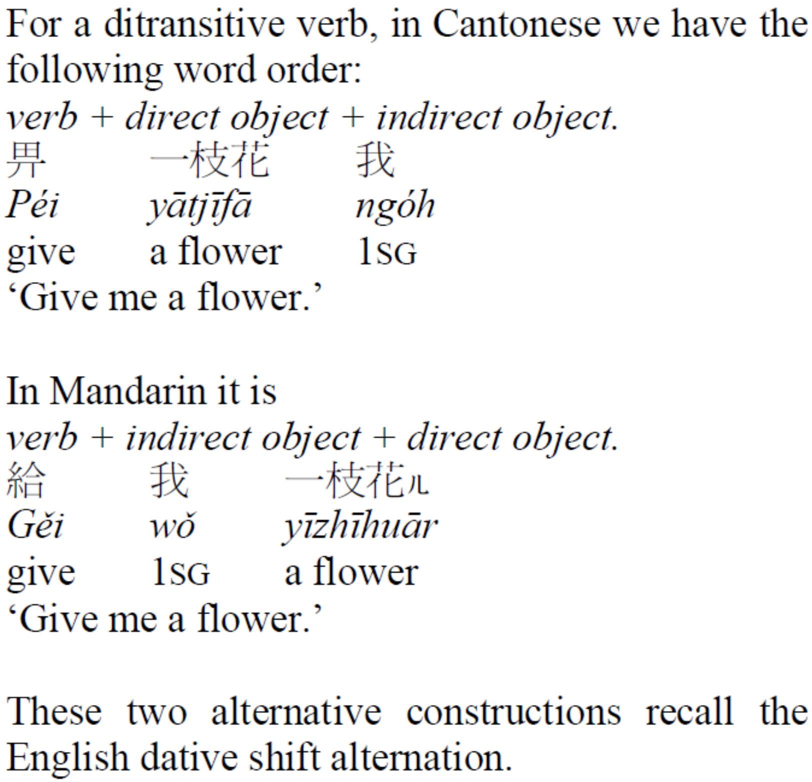
Fisher Test and Specificity

$$\text{Specificity} = \begin{cases} -\log_{10}(p) \\ \log_{10}(1-p) \end{cases}$$

- Cantonese: lower frequency of adverbs
- prominence of Cantonese post-verbal particles
- Mandarin: uses adverb more often
- Mandarin: *zhèngzài* + V
- Cantonese: V-*gán*

Some Interesting Constructions

Double objects



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

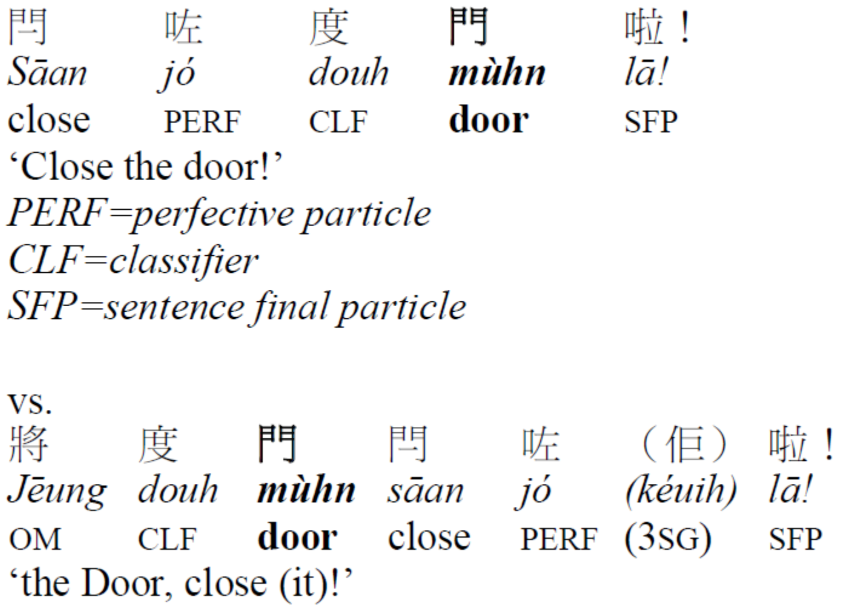
verb + indirect object + direct object.

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

‘Give me a flower.’

These two alternative constructions recall the English dative shift alternation.

Object marker



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.

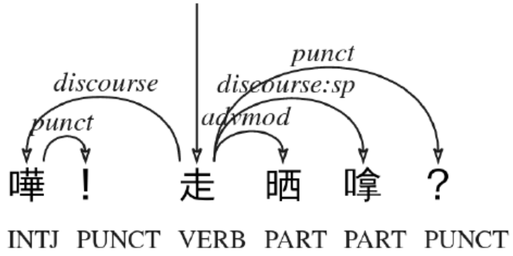
將	度	**門**	閂	咗	（佢）	啦！
Jēung	*douh*	***mùhn***	*sāan*	*jó*	*(kéuih)*	*lā!*
OM	CLF	**door**	close	PERF	(3SG)	SFP

‘the Door, close (it)!’

Some Interesting Constructions

Post-verbal modifiers

Cantonese:

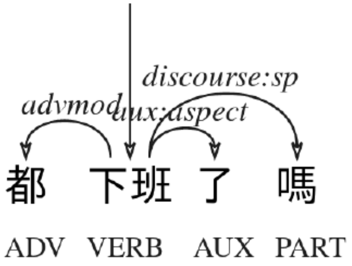


嘩 ！ 走 晒 嗱 ？

INTJ PUNCT VERB PART PART PUNCT

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:



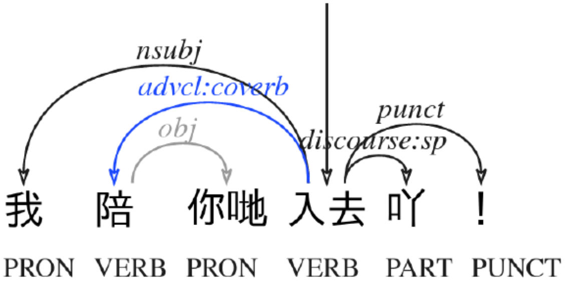
都 下班 了 嗎

ADV VERB AUX PART

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

Coverb constructions

Cantonese:

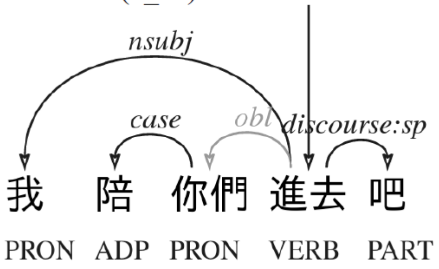


我 陪 你哋 入去 吖 ！

PRON VERB PRON VERB PART PUNCT

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

Mandarin (0_28):

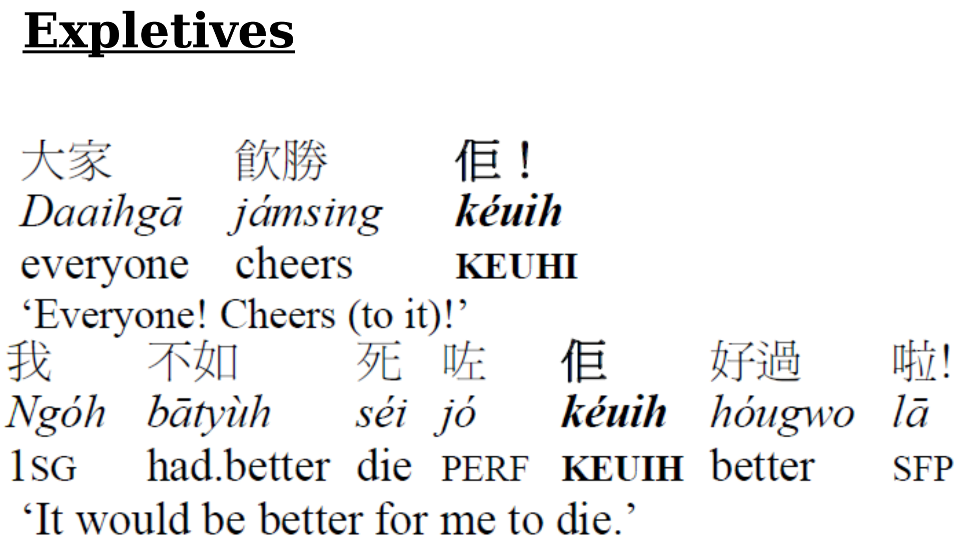


我 陪 你們 進去 吧

PRON ADP PRON VERB PART

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

Some Interesting Constructions



Expletives

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.’