



# Exploring the Role of Sentence-Final Particles in Spoken Cantonese: A Comparative Study Based on the POS-based Dependency Networks

Yixi Chen<sup>1</sup> , Jianwei Yan<sup>1\*</sup> 

<sup>1</sup> Department of Linguistics, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, China.

\* Corresponding author's email: [jwyan@zju.edu.cn](mailto:jwyan@zju.edu.cn)

DOI: [https://doi.org/10.53482/2024\\_57\\_419](https://doi.org/10.53482/2024_57_419)

## ABSTRACT

Sentence-final particles (SFPs) are pervasive in spoken Cantonese for expressing speakers' attitudes. This study explores the global and local features of SFPs in both spoken Cantonese and Mandarin Chinese using two part-of-speech-based (POS-based) dependency networks. Results show that (1) globally, spoken Cantonese and Mandarin Chinese networks exhibit centralization and scale-free properties, reflecting the communication efficiency of human languages. However, spoken Cantonese manifests weaker centralization properties, as demonstrated by the diversity of its edges. Moreover, SFPs in spoken Cantonese have a greater degree and in-degree than those in Mandarin Chinese, indicating a stronger ability to form syntactic connections with other POSs in the language structure. (2) locally, Cantonese SFPs display more extensive mood expression devices, notably differing in PART-NOUN (discourse:sp), PART-ADV and PART-ADJ dependencies compared to Mandarin Chinese. Additionally, specific examples illustrate how Cantonese SFP usage differs from Mandarin Chinese, showcasing their distinct discourse functions. The findings suggest that communication efficiency is a cross-lingual universal, while spoken Cantonese is distinctive in its use of diverse SFPs to express moods. This study may shed new light on adapting the complex network approach to explore the similarities and differences across human languages.

**Keywords:** sentence-final particle, complex network, dependency grammar, part-of-speech, spoken Cantonese

## 1 Introduction

Sentence-final particles (SFPs) in spoken Cantonese, a variety of Chinese belonging to the Yue sub-branch, constitute a rich system widely used in daily communication (Law, 1990; Luke, 1990). They provide a diverse inventory of grammatical devices to express sentence mood (Chor, 2018; W. Leung, 2006; Luke, 1990; Yip, 1994). Although Mandarin Chinese and English also have their own SFPs, those in Cantonese feature greater subtlety and diversity in mood expression. Given the sentence *keoi5 hai2*

*dou6 wan2 je5* ‘they<sup>1</sup> are looking for something’ in Example (1), the attachment of different SFPs in Cantonese conveys various meanings, as shown in Table 1. Details regarding tone numbers can be found in the Appendix.

(1)	佢喺度搵嘢。			
	<i>keoi5</i>	<i>hai2-dou6</i>	<i>wan2</i>	<i>je5</i>
	她/他	在	找	东西
	<i>ta1</i>	<i>zai4</i>	<i>zhao3</i>	<i>dong1-xi</i>
	They	PROG <sup>2</sup>	look for	thing
	‘They are looking for something.’			

Table 1 presents three examples of SFPs (*wo5*, *gwaa3* and *ze1*) attached to Example (1). The attachment of *wo5* expresses speakers’ certainty about the fact that “they are looking for something”. This SFP also serves to emphasize a noteworthy piece of information in certain context (Yip, 1994). *Gwaa3* can be interpreted as a particle expressing the speaker’s uncertainty and speculation about the stated fact. The attachment of *ze1* downplays the importance of the stated fact.

**Table 1:** Examples of SFP attachment in Cantonese.

Attached SFP	Sentence					Description of the SFP
喎 ( <i>wo5</i> )	<i>keoi5</i>	<i>hai2-dou6</i>	<i>wan2</i>	<i>je5</i>	<i>wo5</i>	Certainty about the stated fact
	‘ <u>It is said that</u> they are looking for something.’					
喺 ( <i>gwaa3</i> )	<i>keoi5</i>	<i>hai2-dou6</i>	<i>wan2</i>	<i>je5</i>	<i>gwaa3</i>	Uncertainty about the stated fact
	‘ <u>It is probable that</u> they are looking for something.’					
啫 ( <i>ze1</i> )	<i>keoi5</i>	<i>hai2-dou6</i>	<i>wan2</i>	<i>je5</i>	<i>ze1</i>	Understatement of the stated fact
	‘ <u>(Nothing.)</u> They are <u>just</u> looking for something.’					

Despite the contribution of SFPs to mood expression, they are semantically weak and optional in language usage (Luke, 1990). Moreover, compared to spoken Cantonese, spoken Mandarin Chinese (which is syntactically similar to Cantonese) also uses SFPs to express moods, but these particles are more “general in meaning and broad” (Law, 1990). The richness of SFP in spoken Cantonese may therefore seem unnecessary and may require extra effort in language processing. Law (1990) claimed that the diverse use of SFP in Cantonese might stem from its restrained tonal systems. The restrained system renders Cantonese phonologically freer in pitch variation, resulting from both tones and intonation. In other words, the variation of pitches in Cantonese relies more on its six tones, while in Mandarin Chinese, the variation of pitches results more from intonation since it only has four tones (Law, 1990). As Yau (1980, p.51) stated, “the more a language relies on the use of sentence particles [i.e., SFP in the current study] in expressing sentential connotations, the less significant will be the role played by

<sup>1</sup> *Keoi5* is a gender-neutral third person singular pronoun in Cantonese, which is translated as ‘they’ in the current study.

<sup>2</sup> The list of glossing abbreviations can be seen in <https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf>.

intonation patterns, and vice versa.” Therefore, the differences in tonal systems provide a possible explanation for the richness of SFPs in spoken Cantonese.

Apart from Yau (1980) and Law (1990), there has been little thorough explanation for the pervasiveness of SFPs in spoken Cantonese. If the pervasiveness and richness of SFPs in spoken Cantonese are not trivial, their influence on mood should not be solely contextual, which is dispensable, but also structural. In other words, Cantonese SFPs should have a strong capacity to connect with other language components, rendering Cantonese structurally different from other languages. Therefore, our comparison between two structurally similar languages (i.e., spoken Cantonese and Mandarin Chinese) is valuable in exploring the extent to which SFPs in spoken Cantonese induce differences in syntactic structure.

In addition, focusing solely on specific linguistic examples may fall short of a global view of a language. Human language is a complex system where linguistic components are mutually connected in terms of syntax and their relationships constitute edges in the syntactic network (Gibson et al., 2019; Hawkins, 2004; Liu, 2008). This system can be modeled using the complex network approach based on large-scale corpora data (Liu, 2008; Liu & Xu, 2011; Cong & Liu, 2014). Consistent with previous studies, we focused on these two properties as critical global features and examined the cumulative distributions of degrees to deepen insight into the two networks’ scale-free property and hub nodes. In the syntax network of a language, degree centralization is related to the capacity of central linguistic components (as the hub nodes) to build syntactic connections with others (Chen & Liu, 2016; Ferrer-i-Cancho, 2013; Liu, 2008), while scale-free property manifests the efficiency of syntax to organize linguistic components in use (Liu & Xu, 2011; Chen et al., 2018; Yang & Liu, 2022). Both centralization and scale-free properties are expected to be universal in the language network, as proved by the aforementioned studies. Meanwhile, differences can be found in the comparison of related metrics and the parameters of the cumulative distribution of degrees.

Besides the global investigation, the examination of specific linguistic examples was also combined to illustrate and explain the differences of SFP in language usage. This is because SFP in spoken Cantonese has both intrinsic meaning (Wakefield, 2011) and contextual meaning (Luke, 1990), whose interpretation should thus be embedded in the context. Local investigation enables us to focus on SFP and its related dependencies in the language system and facilitates detailed interpretation within the context. This part of the discussion should contribute to an integrative scope of SFP in spoken Cantonese when extending the line of previous studies.

Aiming to examine the role of SFP in the language system of spoken Cantonese, we integrated both the global features of the language network and the local features of SFP in spoken Cantonese and Mandarin Chinese. To this end, the specific research questions are as follows:

**Question 1:** What similarities and differences can be found in the POS-based dependency networks of spoken Cantonese and Mandarin Chinese?

**Question 2:** How are the differences in SFPs reflected in POS-based dependency networks and specific linguistic examples?

**Question 1** focuses on the global features of spoken Cantonese compared to Mandarin Chinese. In response to Question 1, we built two language networks of spoken Cantonese and spoken Mandarin Chinese based on dependency grammar, where dependencies refer to the binary asymmetric syntactic relation between two words (Hudson, 2007, 2010; Liu, 2009). Notably, we converted the word-based dependencies to POS-based dependencies to explicitly examine the dependencies between SFP and other categories of words. Global features of spoken Cantonese and Mandarin Chinese are discussed in Section 3.1, where Section 3.1.1 focuses on the centralization and scale properties and Section 3.1.2 on the in-degree and out-degree distributions of the two networks.

As indicated by Liu et al. (2010), it is not enough to build an explainable link between the local syntactic structure and the global behavior if we solely focus on quantitative metrics in a linguistic network. A local investigation combined with specific examples is thus necessary. Hence, **Question 2** focuses on the local features of SFPs in spoken Cantonese. We examined SFPs in language usage by delving into specific examples in Section 3.2. We first selected a range of dependencies that manifested discrepancy between two languages in Section 3.2.1 and then looked into three SFP-related dependencies, namely PART-NOUN (discourse:sp), PART-ADV, and PART-ADJ, in Section 3.2.2. Specific examples of SFPs used in the context may offer empirical evidence for its role in mood expression.

## 2 Data and Methods

### 2.1 Data

Data in the current study consists of two paralleled treebanks retrieved from the database of Universal Dependencies (UD)<sup>3</sup>: UD Cantonese HK (UD-CANT,  $N$  of sentences = 1,004,  $N$  of tokens = 13,918) and UD Chinese HK (UD-CHIN,  $N$  of sentences = 1,004,  $N$  of tokens = 9,874) (Wong et al., 2017). The UD framework ensures cross-lingual consistency in the annotation and thus has been well adopted in cross-linguistic comparative and typological studies (Levshina, 2019; Yan & Liu, 2023). The application of the paralleled treebank UD-CHIN not merely provides reliable Mandarin Chinese translations but also facilitates the comparison in terms of global features.

---

<sup>3</sup> UD treebanks are open-sourced and available on <https://universaldependencies.org/>. The version used in the current study is 2.12, released on May 15, 2023.

Texts in both treebanks belonged to the spoken genre, containing subtitles from three films ( $N$  of sentences = 650) and part of the official records of the council discourse ( $N$  of sentences = 354). Each sentence in the UD-CANT is matched with a semantic counterpart in the UD-CHIN, as is shown in Example (2), where *aa3* in Example (2b) was annotated as a sentence-final particle (SFP).

- (2) a. 你在找些什麼?  
*ni3 zai4 zhao3 xie1 shen2-me*  
 you PROG look for CLF WH
- b. 你喺度搵乜嘢呀?  
*nei5 hai2-dou6 wan2 mat1-je5 aa3*  
 you PROG look for WH SFP  
 ‘What are you looking for?’

Sentences in both treebanks were all used in the following data analysis. Both treebanks were manually annotated in terms of POS tags and dependency relations and coded in CoNLL-U files, which are commonly used in dependency annotation, as shown in Table 2. Therein, SFP is annotated as “PART”.

**Table 2:** Dependency structure of Example (2b) 你喺度搵乜嘢呀?

Token	Token Order	Token POS	Head Order	Head POS	DEP
<i>nei5</i>	1	NOUN	3	VERB	nsubj
<i>hai2-dou6</i>	2	ADV	3	VERB	advmod
<i>wan2</i>	3	ROOT	0		root
<i>mat1-je5</i>	4	PRON	3	VERB	obj
<i>aa3</i>	5	PART	3	VERB	discourse:sp

Note: DEP, dependency relation.

Data extraction was conducted in Python 3.10.1. Frequencies of POSs and dependencies in the two treebanks were calculated. In this step, dependencies annotated as “root” (root word) and “punct” (punctuation) were eliminated since our focus was on the dependency relation between POSs. General information on the two treebanks is shown in Table 3.

**Table 3:** General information on UD-CANT and UD-CHIN.

		Type	Frequency		
			<i>Mean</i>	<i>Mdn</i>	<i>SD</i>
UD-CANT	POS	14	799	352	811.62
	DEP	124	82.14	8.5	229.11
UD-CHIN	POS	15	542	301	609.18
	DEP	101	70.58	7	183.88

## 2.2 Methods

Data analysis was conducted in R 4.3.2 (R Core Team, 2023). Packages used are as follows: *igraph* 1.5.1 (Csardi & Nepusz, 2006) and *ggraph* 2.1.0 (Pedersen, 2023) to construct dependency networks and *ggplot2* 3.4.4 (Wickham, 2016) to visualize results.

The basic elements of a dependency network contain nodes representing tokens and directed edges representing dependency relations (Liu, 2008). To construct POS-based dependency networks for the two treebanks, POSs and dependencies in syntactic structures were first converted to nodes and edges, respectively, as shown in Table 4. By doing so, we can focus on the dependencies between SFP and other categories of words (instead of the dependencies between specific words) and more explicitly examine SFP’s role in human language networks. Notably, each edge in the dependency network is directed, pointing from the head POS to the governed token POS.

**Table 4:** Dependency network structure of Example (2b) 你喺度搵乜嘢呀?

Token	Start (head POS)	End (token POS)	Edge (DEP)
<i>nei5</i>	VERB	NOUN	nsubj
<i>hai2-dou6</i>	VERB	ADV	advmod
<i>wan2</i>		VERB	
<i>mat1-je5</i>	VERB	PRON	obj
<i>aa3</i>	VERB	PART	discourse:sp

Note: start, the start of each edge; end, the end of each edge.

An important metric in the dependency network is the degree  $k_i$ , referring to the number of edges that connect the  $i$ -th node with others. In practice, it can be calculated as the number of types of relevant dependencies<sup>4</sup>. The degree of a node can be sub-divided into the in-degree  $k_{in}$  and the out-degree  $k_{out}$ , where the former measures the number of dependencies uniquely directed to the node and the latter from the node. The degree  $k$  of the token *wan2* (VERB) is 4 in the network structure shown in Table 4, whose in-degree is 0 (no dependency is directed *to* VERB) while the out-degree is 4 (four dependencies are directed *from* VERB respectively to NOUN, ADV, PRON and PART).

Regarding the centralization property, an important metric is the density  $\rho$ . It is defined as the ratio between the observed number of edges  $m$  in the network and its theoretical maximum  $n(n - 1)$  ( $n$  refers to the number of nodes) in a directed network<sup>5</sup>. The formula is shown in (1).

$$\rho = \frac{m}{n(n - 1)} \quad (1)$$

<sup>4</sup> Notably, the combination of the same pair of POSs can be different, for example the dependency relation of POSs “NOUN” (dependent) and “VERB” (head) can be annotated as “nsubj” or “obj”.

<sup>5</sup> In an undirected network with  $n$  nodes, the theoretical maximum is  $\frac{n(n-1)}{2}$  where each pair of nodes maximally allow one edge between them. Whereas, in a directed network, two edges are allowed between a pair of nodes (A to B, B to A).

Based on the density, degree centralization  $NC$  was calculated, which reflected the importance of hub nodes for the whole network. It is calculated as (2) (Butts, 2006; Freeman, 1978):

$$NC = \frac{n}{n-1} \left( \frac{k_{max}}{n-1} - \rho \right) \quad (2)$$

The scale-free property is related to the cumulative distribution of degrees (Barabási & Albert, 1999; Ferrer-i-Cancho & Solé, 2003; Newman, 2003, 2010). This property indicates that only a small number of nodes have high degrees in the network, conforming to the principle of least effort (Zipf, 1949). It can be formulated as:

$$P(k) \sim k^{-\gamma} \quad (3)$$

In (3), the degree distribution  $P(k)$  is the cumulative probability that a random node has a degree greater than or equal to  $k$ .  $\gamma$  is the parameter, which is expected to be negative and indicates the decreasing trend of  $P(k)$  as the cumulative sum of  $k$  increases.

### 3 Results and Discussion

#### 3.1 Global features of spoken Cantonese and spoken Mandarin Chinese

In response to the first question, we looked into the global features of the two languages of interest. To illustrate, we first constructed the POS-based dependency networks according to the paralleled treebanks UD-CANT and UD-CHIN to testify to centralization and scale-free properties in Section 3.1.1. These two properties are expected to be similar to both spoken Cantonese and Mandarin Chinese. Then, we investigated their cumulative distributions of degrees in Section 3.1.2, including both in-degrees and out-degrees, to examine the differences between these two languages.

##### 3.1.1 Centralization and scale-free properties in POS-based dependency networks

Two POS-based dependency networks based on UD-CANT and UD-CHIN are presented in Figure 1. Nodes in the network are POSs and edges are the corresponding dependencies. The size and alpha of each label in each figure represent the degree of the POS, and the width of each edge is related to the frequency of the dependency, which is also annotated on the edge. The two networks show the prominent nodes in spoken Cantonese and spoken Mandarin Chinese. It is found that NOUN (noun)<sup>6</sup>, VERB (verb), ADV (adverbial), ADJ (adjective), PROP (proper noun) and PRON (pronoun) are highly frequent in both languages. These POSs also serve as the hub nodes, bearing rich connections with other

<sup>6</sup> The description of POS tags is based on <https://universaldependencies.org/u/pos/index.html>.

nodes in the dependency network. One noteworthy point is that PART (particle), representing SFP, serves as the hub node only in spoken Cantonese.

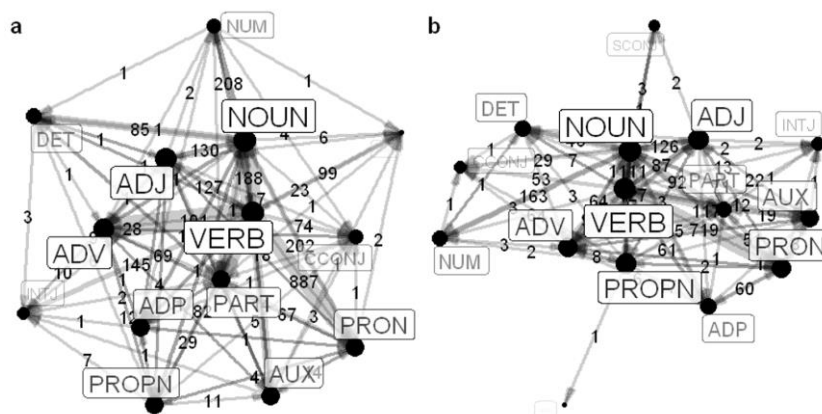


Figure 1: POS-based dependency networks of UD-CANT (a) and UD-CHIN (b).

Network density  $\rho$  and degree centralization  $NC$  were then calculated to investigate the centralization property of the two networks. Results are shown in Table 5. Greater  $\rho$  indicates that the network of spoken Cantonese manifests stronger centralization properties. In other words, the nodes therein have a stronger capability of connecting with others than their counterparts in spoken Mandarin Chinese. The prominence of PART aforementioned is equally captured by its greater degree in spoken Cantonese ( $k_{PART} = 20$ ) than that in spoken Mandarin ( $k_{PART} = 11$ ). It supports the richness of SFP as a peculiarity in spoken Cantonese.

Table 5: General information on the two POS-based dependency networks.

	<i>n</i> of nodes	<i>n</i> of edges	<i>k</i>	<i>k</i> <sub>in</sub>	<i>k</i> <sub>out</sub>	$\rho$	<i>NC</i>
UD-CANT	14	124	18.64	8.86	9.79	.86	1.42
UD-CHIN	15	101	14.33	6.80	7.53	.48	1.47

Note: *k*, the mean degree; *k*<sub>in</sub>, the mean in-degree, *k*<sub>out</sub>, the mean out-degree.

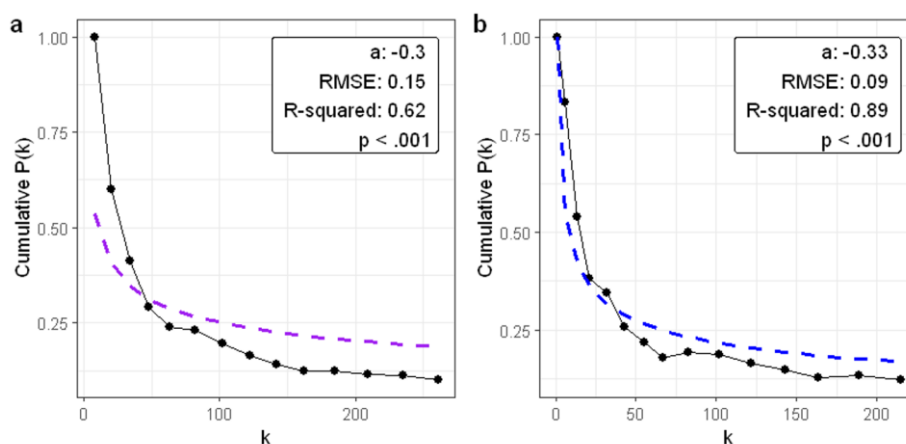


Figure 2: Cumulative distributions of degrees based on UD-CANT (a, purple dashed fitted curve) and UD-CHIN (b, blue dashed fitted curve), where the function of the degree distribution is  $P(k) \sim k^a$  ( $a < 0$ ).

The scale-free property is validated by the goodness-of-fit of the degree distribution. Figure 2 presents the cumulative degree distributions based on UD-CANT and UD-CHIN. The y-axis is the cumulative probability of each node, and the x-axis is the cumulative sum of  $k$ . The black curves represent the observed frequencies, while the purple and blue curves represent the predicted values. The degree distributions of POSs in UD-CANT (RMSE = 0.15,  $R^2 = .62$ ,  $p < .001$ ) and UD-CHIN (RMSE = 0.09,  $R^2 = .89$ ,  $p < .001$ ) both conformed to the power-law distribution.

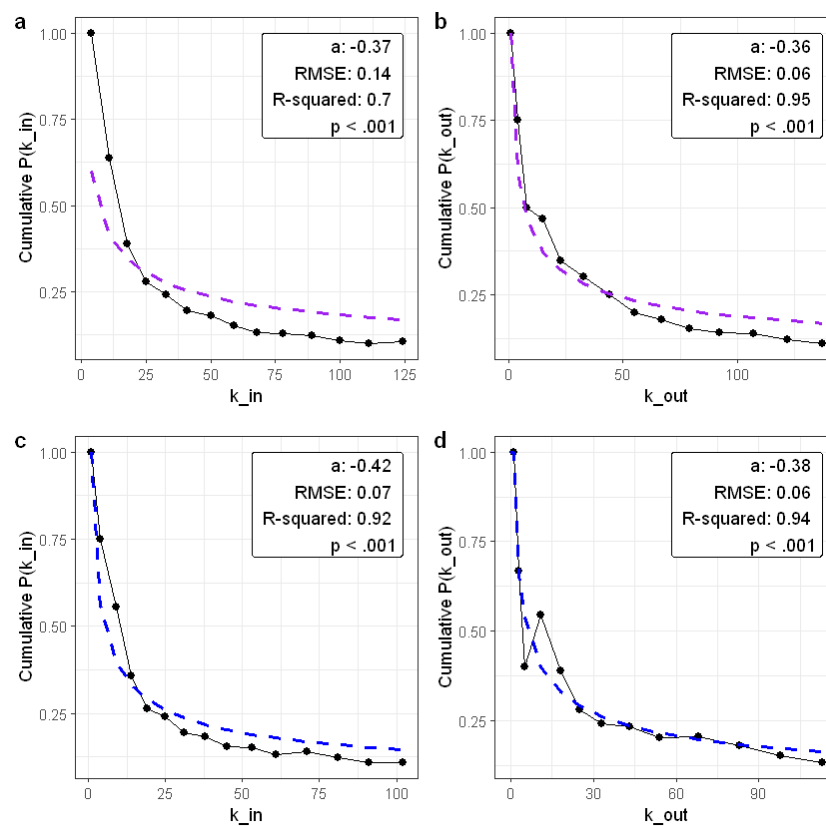
Both Figure 1 and Figure 2 suggest that only a small number of hub nodes are structurally important and connected to a large number of nodes in the two dependency networks. Consistent with previous studies (Liu, 2008; Liu & Hu, 2008; Yang & Liu, 2022), our results indicate that the scale-free property and centralization property are common features in both spoken Cantonese and spoken Mandarin Chinese, agreeing with the principle of least effort in the language system.

### 3.1.2 In-degree and out-degree distributions in POS-based dependency networks

To explore the differences between the two languages, we further investigated the in- and out-degree distributions of POS-based dependencies. The degree of a node relates to both the frequency of the node itself and its capability to combine with different nodes and carry information on different sentence constituents, which refers to its valency (Čech et al., 2011). However, valency is not equivalent to the degree of nodes. Valency is related to the out-degree of a node (rather than the in-degree). In this section, the degrees of each node were adjusted into in-degrees and out-degrees. The in-degree and out-degree distributions based on UD-CANT and UD-CHIN are shown in Figure 3.

Cumulative distributions of in-degrees and out-degrees based on the two treebanks also satisfactorily fitted the power-law distribution, validating the scale-free property. Out-degree distributions yielded higher goodness of fit ( $R^2 = .95$ ,  $p < .001$ ;  $R^2 = .94$ ,  $p < .001$ ) compared to in-degree distributions ( $R^2 = .70$ ,  $p < .001$ ;  $R^2 = .92$ ,  $p < .001$ ). Notably, the fitted curves of spoken Mandarin Chinese (Figure 3c and Figure 3d) are skewer than those of spoken Cantonese (Figure 3a and Figure 3b), which was supported by the parameter  $a$ .

Results validate the greater differences among nodes in spoken Mandarin Chinese in terms of both in-degrees and out-degrees, which aligns with its stronger centralization property. The discrepancy of the two languages stems from their syntax. Yang and Liu (2022) stated that the role of syntax might “widen[s] the differences among degrees”. Although spoken Cantonese allows more diverse dependencies, the organization of sentences majorly relies on a small number of POSs (such as VERB) and a limited range of dependencies. This renders the remaining majority of dependencies less frequent in language use. The remaining majority provides a diverse choice of dependencies for language users to combine different categories of words in language production, manifesting the flexibility and variability of spoken Cantonese.



**Figure 3:** Cumulative distributions of in-degrees (a, c) and out-degrees (b, d) based on UD-CANT (purple dashed fitted curve) and UD-CH (blue dashed fitted curve), where the function of the degree distribution is  $P(k) \sim k^a$  ( $a < 0$ ).

While the degree distribution provided an overview to compare the two languages, their differences were also reflected in the rank of POS. In other words, a POS may be preferred in one language but disfavored in another. We summarized the ranked distribution of the top five POSs in spoken Cantonese and spoken Mandarin Chinese in Table 6. The ranked distribution in terms of degrees and out-degrees does not show much difference between the two languages, demonstrating their similarity in syntax. Specifically, VERB showed high degree and out-degree in the two languages, validating its strong capability or valency in the language structure (Čech et al., 2011).

**Table 6:** Ranked distribution of POSs in terms of degrees, in-degrees and out-degrees.

	UD-CANT (degree)	UD-CHIN (degree)	UD-CANT (in-degree)	UD-CHIN (in-degree)	UD-CANT (out-degree)	UD-CHIN (out-degree)
1	NOUN (26)	NOUN (26)	PART (13)	NOUN (11)	ADJ (15)	PROPN (15)
2	VERB (26)	VERB (25)	ADV (11)	PART (10)	NOUN (15)	VERB (15)
3	ADJ (24)	ADJ (21)	NOUN (11)	ADV (10)	VERB (15)	NOUN (15)
4	ADV (23)	PROPN (21)	VERB (11)	VERB (10)	PROPN (13)	ADJ (14)
5	ADP (20)	ADV (20)	CCONJ (10)	AUX (8)	PRON (12)	PRON (11)
...	...	...	...	...	...	...

Moreover, some major differences can be found in the ranked distribution of in-degrees. PART in spoken Cantonese yielded a higher in-degree ( $k_{in} = 13$ ) than it is in spoken Mandarin Chinese ( $k_{out} = 10$ ), and CCONJ (coordinating conjunction) only showed a high in-degree in spoken Cantonese. The high in-degree and low out-degree of PART does not mean that its role is insignificant in spoken Cantonese. Low out-degree was found advantageous to shorten the average path length in the network and lessen language users' efforts in language production, which is supported by the study of Chen and Liu (2016) on Mandarin Chinese function words (*de*, *le*, and *zhe*). Such an explanation may be analogizable to SFP in spoken Cantonese: the richness of SFP provides a diverse choice of dependencies while its capacity to be governed by other POSs allows “shortcuts” in the network to ensure communication efficiency.

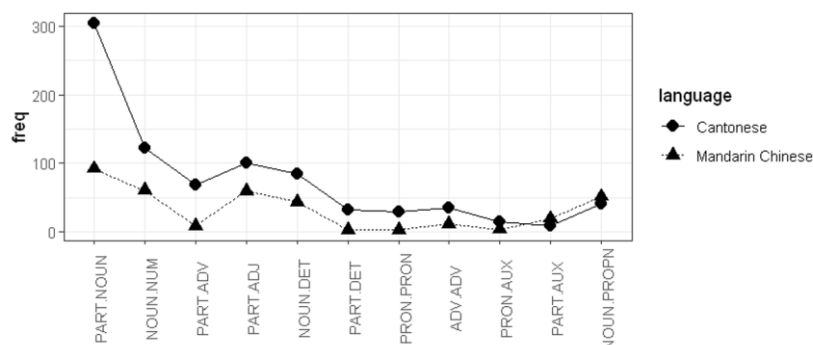
### 3.2 Local features of SFP in spoken Cantonese and Mandarin Chinese

The previous section validated the universal presence of centralization and scale-free property. Differences were also found. The network of spoken Cantonese reflects weaker centralization properties, suggesting more flexibility and diversity in its syntactic structure. In addition, PART is only found prominent in spoken Cantonese for its high degree and in-degree, which is in line with the richness of particles discussed in previous studies (Chor, 2018; Law, 1990; Luke, 1990; Sybesma & Li, 2007). We first summarized the local features of SFP and highlighted three dependencies (i.e., PART-NOUN, PART-ADV, and PART-ADJ.) in the networks in Section 3.2.1. The interpretation of specific linguistic examples is then presented in Section 3.2.2 to provide empirical evidence of the differences between the two languages in SFP.

#### 3.2.1 SFP in POS-based dependency networks

As mentioned, a (type of) dependency may be highly frequent in one language but less preferred in another. These dependencies were the focus of this section since they can manifest the structural discrepancies between two languages well. Here, the discrepancy in use was measured by the cross-

treebank differences in frequency and frequency rank. Hence, we selected those dependencies with (1) cross-treebank differences of frequencies greater than ten and (2) cross-treebank differences of ranks greater than five. Results are illustrated in Figure 4. Among the 11 dependencies selected, PART serves as the governed POS in five dependencies. This means that the structural discrepancy of PART between the two languages majorly originates from dependencies where it serves as the dependent rather than the head.



**Figure 4:** Comparison of the frequencies of POS-based dependencies between UD-CANT and UD-CHIN.

Table 7 summarizes the frequencies of dependencies related to PART in the two treebanks, where a positive value of “Difference” means the corresponding dependency is more frequent in UD-CANT. It can be seen that PART-NOUN (discourse:sp) and PART-ADV are highly frequent in spoken Cantonese while sparse in spoken Mandarin Chinese. PART-ADV and PART-DET dependencies are only found in spoken Cantonese. In contrast, PART-AUX is unique to spoken Mandarin Chinese in the comparison.

**Table 7:** Cross-treebank frequencies of dependencies related to PART.

Token POS	Head POS	DEP	Freq. (UD-CANT)	Freq. (UD-CHIN)	Difference
PART	NOUN	discourse:sp	189	17	172
		case	101	70	31
PART	ADV	discourse:sp	58	0	58
PART	ADJ	discourse:sp	57	25	32
PART	DET	case	30	0	30
PART	AUX	discourse:sp	0	19	-19

Note: Freq., frequency; in the current table, only dependencies with frequencies greater than ten were retained.

In sum, the comparison of cross-treebank frequency offers a more detailed illustration of SFP (represented by PART here) in spoken Cantonese and Mandarin Chinese. Dependencies with PART as the dependent are more frequent in spoken Cantonese, which is in line with its prominence as the hub node and high in-degree in the POS-based network. Besides, these dependencies are majorly related to discourse function (discourse:sp), demonstrating SFP’s role in mood expression.

### 3.2.2 SFP in specific linguistic examples

The previous section quantitatively identified the dependencies showing differences between these two languages. However, investigation solely based on quantitative metrics is not explainable enough to identify the differences in SFPs used in specific contexts. We thus turned to specific language examples to further understand the discrepancy between the local features of spoken Cantonese and Mandarin Chinese. This section mainly focused on three types of dependencies: PART-NOUN (discourse:sp), PART-ADV, and PART-ADJ.

- (3) a. 喂，你如果係選嘅話，一陣半個鐘頭之內，十分鐘之內添，佢就已經係主席喇嘞。
- |                  |                   |                  |                    |                  |                         |
|------------------|-------------------|------------------|--------------------|------------------|-------------------------|
| <i>wai3</i>      | <i>nei5</i>       | <i>jyu4-gwo2</i> | <i>hai6</i>        | <i>syun2</i>     | <i>ge3-waa6</i>         |
| INTJ             | you               | if               | BE                 | elect            | CONJ                    |
| <i>jat1-zan6</i> | <i>bun3</i>       | <i>go3</i>       | <i>zung1-tau4</i>  | <i>zil-noi6</i>  |                         |
| later            | half              | CLF              | hour               | within           |                         |
| <i>sap6</i>      | <i>fan1-zung1</i> | <i>zil-noi6</i>  | <u><i>tim1</i></u> |                  |                         |
| ten              | minutes           | within           | <u>SFP</u>         |                  |                         |
| <i>keoi5</i>     | <i>zau6</i>       | <i>ji5-ging1</i> | <i>hai6</i>        | <i>zyu2-zik6</i> | <u><i>gaa4-laa3</i></u> |
| they             | ADV               | already          | BE                 | president        | <u>SFP</u>              |
- b. 但是，如果要進行選舉的話，有人將於稍後半小時，或十分鐘後便會當選主席。
- |                   |                    |                    |                   |                   |                |
|-------------------|--------------------|--------------------|-------------------|-------------------|----------------|
| <i>dan4-shi4</i>  | <i>ru2-guo3</i>    | <i>yao4</i>        | <i>jin4-xing2</i> | <i>xuan3-ju3</i>  | <i>de-hua4</i> |
| but               | if                 | want               | PROG              | election          | CONJ           |
| <i>you3</i>       | <i>ren2</i>        | <i>jiang1</i>      | <i>yu2</i>        | <i>shao1-hou4</i> |                |
| EXIST             | someone            | will               | in                | later             |                |
| <i>huo4</i>       | <i>shi2</i>        | <i>fen1-zhong1</i> | <i>hou4</i>       |                   |                |
| or                | ten                | minutes            | later             |                   |                |
| <i>bian4-hui4</i> | <i>dang1-xuan3</i> | <i>zhu3-xi2</i>    |                   |                   |                |
| will              | elect-PASS         | president          |                   |                   |                |
- ‘If you want to join in the election, within half an hour or ten minutes, you will be the president.’

Example (3a) contains two PART-NOUN (discourse:sp) dependencies: SFP *tim1* is dependent on *fan1-zung1* ‘minutes’ and SFP *gaa4-laa3* on *zyu2-zik6* ‘president’. The SFP *tim1* is semantically approximate to “too”, “also” or “even” in English (Lee & Pan, 2010) and suggests a greater degree of the element to which it is attached (*sap6 fan1-zung1 zil-noi6* ‘within ten minutes’) than its counterpart (*jat1-zan6 bun3 go3 zung1-tau4 zil-noi6* ‘within half an hour’) in terms of the imprudence to make the decision. The SFP *gaa4-laa3* is usually used to express certainty (Chor, 2018) or seek mutual agreement or common assessment (Luke, 1990). The use of SFP attaches the speaker’s attitude to the fact related to the head noun and expresses the speaker’s irony for the election process.

- (4) a. 中間唔要有空白呀，知唔知呀？  
*zung1-gaan3 m4 jiu3 jau6 hung1-baak6 aal*  
 middle NOT AUX EXIST blank SFP  
*zi1 m4 zi1 aa1*  
 know NOT know SFP
- b. 中間不要有空白，明白嗎？  
*zhong1-jian1 bu2 yao4 you3 kong4-bai2*  
 middle NOT AUX EXIST blank  
*ming2-bai2 ma*  
 know SFP  
 ‘No blank in the middle, understand?’

Example (4a) presents one PART-ADV dependency, namely SFP *aal* and the negative ADV *m4* ‘not’. The SFP *aal* has been described as a particle expressing certainty or doubtlessness (C.-S. Leung, 1992), serving as a softener. In an interrogative sentence as in Example (4a), *aal* softens the negation *m4*, which enables the speaker to lessen the face threat in communication. Without the softener *aal*, Mandarin Chinese and English translations both seem to carry the imperative mood.

- (5) a. 係呀，好煩㗎，佢。  
*hai6 aal hou2 faan4 gaa4 keoi5*  
 BE SFP very annoying SFP she
- b. 對，她有點煩人。  
*dui4 tal you3-dian3 fan2-ren2.*  
 yes she a bit annoying  
 ‘Yes, she is annoying indeed.’

The PART-ADJ dependency between SFP *gaa4* and *faan4* ‘annoying’ in Example (5a) manifests the role of *gaa4* to express determination and assertion (Yip, 1994), whose English translation can be “indeed”. The mood of assertion is diminished in Example (5b) because of the lack of its equivalent in spoken Mandarin Chinese.

As for the remaining three types of dependencies, most of them stemmed from transcription bias. The PART-DET dependency in Example (6a) refers to the dependency *ge3* and *so2-jau6* ‘all’, where the PART *ge3* is not an SFP. In addition, *suo3-you3 tong2-shi4* ‘all the colleagues’ in Example (6b) can be rewritten as *suo3-you3 de tong2-shi4* (*de* is equivalent to *ge3* in Cantonese when expressing the case) with little deviation from the original meaning. A manual review of corpora found that most cases of PART-DET dependency in UD-CANT belong to the current situation. Therefore, it should not be attributed to the structural distinction between languages. Such transcription bias can be analogized to the distribution of PART-NOUN (case) dependencies.

- (6) a. 而家再請所有嘅同事呢，返返去自己原來嘅位置個度。  
*ji4-gaa1 zoi3 cing2 so2-jau6 ge3 tung4-si6 ne1*  
 now again ask all PART colleague SFP  
*fann2 fann2 heoi3 zi6-gei2 jyun5-loi4 ge3*  
 return back go own original PART  
*wai6-zi3 go2-dok6*  
 seat there
- b. 我現在再請所有同事返回自己的座位。  
*wo3 xian4-zai4 zai4 qing3 suo3-you3 tong2-shi4*  
 I now again ask all colleague  
*fan3-hui2 zi4-ji3 de zuo4-wei4*  
 return self PART seat  
 ‘Now, I ask all the colleagues to return to your own seats.’

Another distinction caused by transcription bias is PART-AUX dependency, where the majority of cases in UD-CHIN are the use of *de* to express confirmation (*shi4 de* ‘yes’). Their equivalents in UD-CANT are *hai6 aal* ‘yes’ as is shown in Example (5a), with *hai6* ‘be’ labeled as VERB. The three types of dependencies thus cannot reflect the structural differences of SFP between spoken Cantonese and spoken Mandarin Chinese.

Our discussion of SFP in specific linguistic examples is consistent with previous studies. Combining specific examples of SFP in the treebank, this section illustrates how SFP’s use in spoken Cantonese differs from spoken Mandarin Chinese and how these differences relate to their specific discourse function in communication.

## 4 Conclusion

This study investigated SFPs in spoken Cantonese from a complex network perspective based on two paralleled universal dependency treebanks, namely UD-CANT and UD-CHIN. The perspective from the POS-based dependency network contributes to the global understanding of SFPs’ features in the language structure and provides quantitative evidence for the differences in spoken Cantonese in terms of mood expression.

In terms of global features, POS-based dependency networks of both spoken Cantonese and Mandarin Chinese exhibit centralization and scale-free property properties, in accord with the principle of the least effort. However, spoken Cantonese manifests weaker centralization property, which is demonstrated by the diversity of its edges. Moreover, SFPs show greater degrees and in-degrees in spoken Cantonese than in spoken Mandarin Chinese, indicating its strong capability of building syntactic connections with other POSs in language structure. Further investigation of local features shows that the discrepancy between the two languages lies in PART-NOUN (discourse:sp), PART-ADV and PART-

ADJ dependencies, where SFPs in spoken Cantonese offer an extensive range of devices to express moods in specific usages.

Consistent with previous studies, the current research contributes solid evidence for SFPs as a structurally important component in spoken Cantonese, which is also a feature distinguishable from spoken Mandarin Chinese. By modeling the syntactic system of spoken Cantonese and Mandarin Chinese in the POS-based dependency network, we prove that communication efficiency is a cross-lingual universal, while differences are also shown to adapt to communication needs, such as the expression of speakers' moods in the current case. The pervasiveness of SFPs in spoken Cantonese thus enables different types of dependencies and diversifies the degrees in the dependency network, rendering spoken Cantonese more flexible in expressing speakers' moods. This study also proves the efficacy of the complex network approach in comparing similar languages. Our investigation of both the global features in the dependency network and the local features in linguistic examples constitutes an integrative insight into spoken Cantonese and Mandarin Chinese. Such a method can be used in future studies to examine human language from a comparative perspective. Future studies can focus on more network properties to understand human language systems and their underlying mechanisms to ensure communication efficiency.

## Acknowledgements

This work was supported by the National Social Science Foundation of China (24CYY064).

## References

- Barabási, A.-L., Albert, R.** (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), pp. 509–512. <https://doi.org/10.1126/science.286.5439.509>
- Butts, C. T.** (2006). Exact bounds for degree centralization. *Social Networks*, 28(4), pp. 283–296. <https://doi.org/10.1016/j.socnet.2005.07.003>
- Čech, R., Mačutek, J., & Žabokrtský, Z.** (2011). The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and Its Applications*, 390(20), pp. 3614–3623. <https://doi.org/10.1016/j.physa.2011.05.027>
- Chen, X., Liu, H.** (2016). Function Nodes in Chinese Syntactic Networks. In: Mehler, A., Lücking, A., Banisch, S., Blanchard, P., Job, B. (Eds.). *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pp. 187–201. Springer. [https://doi.org/10.1007/978-3-662-47238-5\\_9](https://doi.org/10.1007/978-3-662-47238-5_9)
- Chen, H., Chen, X., Liu, H.** (2018). How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. *PLoS ONE*, 13(2), e0192545. <https://doi.org/10.1371/journal.pone.0192545>

- Chor, W.** (2018). Sentence final particles as epistemic modulators in Cantonese conversations: A discourse-pragmatic perspective. *Journal of Pragmatics*, 129, pp. 34–47. <https://doi.org/10.1016/j.pragma.2018.03.008>
- Cong, J., Liu, H.** (2014). Approaching human language with complex networks. *Physics of life reviews*, 11(4), pp. 598–618. <https://doi.org/10.1016/j.plrev.2014.04.004s>
- Csardi, G., Nepusz, T.** (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.
- Ferrer-i-Cancho, R.** (2013). *Hubiness, length, crossings and their relationships in dependency trees*. 22.
- Ferrer-i-Cancho, R., Solé, R. V.** (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), pp. 788–791. <https://doi.org/10.1073/pnas.0335980100>
- Freeman, L. C.** (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), pp. 215–239. [https://doi.org/10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7)
- Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., Levy, R.** (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), pp. 389–407. <https://doi.org/10.1016/j.tics.2019.02.003>
- Hawkins, J. A.** (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.
- Hudson, R. A.** (2007). *Language networks: The new word grammar*. Oxford University Press.
- Hudson, R. A.** (2010). *An Introduction to Word Grammar (Cambridge Textbooks in Linguistics)*. Cambridge University Press.
- Law, S.-P.** (1990). *The syntax and phonology of Cantonese sentence-final particles*. Boston University.
- Lee, P. P.-L., Pan, H.-H.** (2010). The landscape of additive particles-with special reference to the Cantonese sentence-final particle *tim*. *LINGUA*, 120(7), pp. 1777–1804. <https://doi.org/10.1016/j.lingua.2009.12.001>
- Leung, C.-S.** (1992). A study of the utterance particles in Cantonese as spoken in Hong Kong. *MPhil. Hong Kong Polytechnic*.
- Leung, W.** (2006). *On the synchrony and diachrony of sentence-final particles: The case of wo in Cantonese* (pp. 991017770249703414, b3622358x) [Doctor of Philosophy, The University of Hong Kong]. [https://doi.org/10.5353/th\\_b3622358](https://doi.org/10.5353/th_b3622358)
- Levshina, N.** (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3), pp. 533-572. <https://doi.org/10.1515/lingty-2019-0025>
- Liu, H.** (2008). The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics and Its Applications*, 387(12), pp. 3048–3058. <https://doi.org/10.1016/j.physa.2008.01.069>
- Liu H.** (2009). *Dependency grammar from theory to practice*. Science Press.
- Liu, H., Hu, F.** (2008). What role does syntax play in a language network? *Europhysics Letters*, 83(1), 18002. <https://doi.org/10.1209/0295-5075/83/18002>

- Liu, H., Xu, C.** (2011). Can syntactic networks indicate morphological complexity of a language? *Europhysics Letters*, 93(2), 28005. <https://doi.org/10.1209/0295-5075/93/28005>
- Liu, H., Zhao, Y., Huang, W.** (2010). How do Local Syntactic Structures Influence Global Properties in Language Networks? *Glottometrics*, 20, pp. 38-58.
- Luke, K. K.** (1990). *Utterance Particles in Cantonese Conversation* (Vol. 9). John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.9>
- Newman, M.** (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2), 167–256. <https://doi.org/10.1137/S003614450342480>
- Newman, M.** (2010). *Networks*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199206650.001.0001>
- Pedersen, T. L.** (2023). *ggraph: An implementation of grammar of graphics for graphs and networks* [Manual]. <https://ggraph.data-imaginist.com>
- R Core Team.** (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Sybesma, R., & Li, B.** (2007). The dissection and structural mapping of Cantonese sentence final particles. *Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Générale*, 117(10), pp. 1739–1783.
- Wakefield, J. C.** (2011). Disentangling the meanings of two Cantonese evidential particles. *Chinese Language and Discourse. An International and Interdisciplinary Journal*, 2(2), pp. 250–293. <https://doi.org/10.1075/cld.2.2.05wak>
- Wickham, H.** (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wong, T., Gerdes, K., Leung, H., & Lee, J.** (2017). Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In: S. Montemagni & J. Nivre (Eds.). *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 266–275. Linköping University Electronic Press. <https://aclanthology.org/W17-6530>
- Yan, J., & Liu, H.** (2023). Basic word order typology revisited: A cross-linguistic quantitative study based on UD and WALs. *Linguistics Vanguard*, 9(1), pp. 73-85. <https://doi.org/10.1515/lingvan-2021-0001>
- Yang, M., & Liu, H.** (2022). The role of syntax in the formation of scale-free language networks. *Europhysics Letters*, 139(6), 61002. <https://doi.org/10.1209/0295-5075/ac8bf2>
- Yau, S.** (1980). Sentential connotations in Cantonese. *Fangyan*, 1, pp. 35–52.
- Yip, S. M., Virginia.** (1994). *Cantonese: A Comprehensive Grammar*. Routledge. <https://doi.org/10.4324/9780203420843>
- Zipf, G. K.** (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

## Appendix: Tones in Cantonese and Mandarin Chinese

Cantonese			Mandarin Chinese		
Tone number	Pitch contour (Five level tone mark)	Example	Tone number	Pitch contour (Five level tone mark)	Example
1	High-level (55)	詩 ( <i>si1</i> )	1	Level (55)	詩 ( <i>shi1</i> )
2	High-rising (25)	史 ( <i>si2</i> )	2	Rising (35)	時 ( <i>shi2</i> )
3	Mid-level (33)	試 ( <i>si3</i> )	3	Dipping (214)	史 ( <i>shi3</i> )
4	Low-falling (21)	時 ( <i>si4</i> )	4	Falling (51)	是 ( <i>shi4</i> )
5	Low-rising (23)	市 ( <i>si5</i> )			
6	Low-level (22)	是 ( <i>si6</i> )			