Speaker-oriented versus listener-oriented production: The case of classifier specificity in Taiwan Mandarin and Chinese Mandarin

Previous studies have shown that linguistic variation is sensitive to information-theoretic (Shannon 1948) constraints: more specific items are used in more unpredictable/informative contexts. This supports listener-oriented production as speakers optimize for specificity to mitigate unpredictability and ensure information transmission. Speaker-oriented production based on information-theoretic effects, however, is also attested, where the speaker chooses more available, rather than specific, items in unpredictable/informative contexts. For example, in Zhan & Levy (2018), the production of general versus specific noun classifiers (i.e., classifier specificity; henceforth CS; cf. 1) in Chinese Mandarin (CM) supported this availability-based account: when the following noun had a higher contextual surprisal, the general classifier *ge* 個, presumably more accessible, was used more than the specific classifiers. Past studies, however, often overlook the possibility of both mechanisms being present and interacting with each other. This study revisits the case of CS in Taiwan Mandarin (TM) and CM, examining its interaction with the following nouns' contextual surprisal/frequency. Using a large-scale corpus of natural speech production, results show that both mechanisms are present and interact with each other in Mandarin classifier production. Additionally, such mechanisms are subject to dialectal differences.

Methodology Data We built a corpus from 2,841 YouTube videos created by Taiwanese and Chinese creators. Classifier-noun pairs were extracted using Stanford CoreNLP (Manning *et al*. 2014). CS was calculated as the mean absolute pointwise mutual information of a classifier with its co-occurring nouns in the corpus (cf. Fig. 1). The noun's contextual surprisal was calculated through the probability distribution extracted from LLAMA-2-7B. Noun frequency was counted directly from the corpus. Statistical analysis Linear mixed-effects regression was used to model CS as a function of noun surprisal/frequency and dialect (TM vs. CM). Additionally, two types of classifiers were identified through elbow analysis: general classifiers (the two most general classifiers, *ge* and *jian* 件) and specific classifiers (others). This was included as a control variable.

Results & discussion Our results showed that noun contextual surprisal had a negative effect on CS. This supports speaker-oriented availability-based production: when the following noun was more unpredictable given context, speakers chose more general/accessible classifiers. Alternatively, noun frequency also had a negative effect on CS. This suggests listener-oriented production: when the noun was less frequent/more informative generally, the more specific classifiers were used. Specifically, a positive interaction was found between noun surprisal and frequency on CS. This suggests competition between the two mechanisms: when the noun frequency is higher, the corresponding specific classifiers are likely more readily accessible. In this case, surprisal can exert a more positive/listener-oriented constraint on CS. Lastly, dialectal differences were found. TM speakers were more positively constrained by surprisal—that is, less speaker-oriented—than CM speakers, and they were even more so when the noun frequency was higher.

Our study demonstrates the presence of both speaker-oriented and listener-oriented production in Mandarin classifier choice. Specifically, the interaction between noun contextual surprisal and noun frequency on classifier specificity suggests crucial interaction between the two mechanisms. The dialectal differences found also suggest that such mechanisms are language-specific.

(1) Examples of more general versus more specific classifiers in Mandarin

A general classifier *ge* 個:	一	個	桌子。	vs.	一	個	背包。
	Yi	*ge*	*zhuo.zi.*		*Yi*	*ge*	*bei.bao.*
	one	CL:GENERAL	table		one	CL:GENERAL	backpack
	“a table”				“a backpack”		
A more specific classifier *zhang* 張:	一	張	桌子。	vs.	一	*張	背包。
	Yi	*zhang*	*zhuo.zi.*	vs.	*Yi*	**zhang*	*bei.bao.*
	one	CL:SURFACE	table		one	CL:SURFACE	backpack
	“a table”				Intended: “a backpack”		

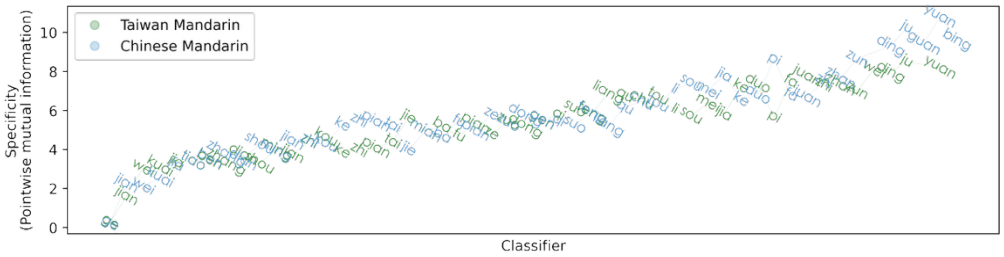

Figure 1. Calculated specificities of the classifiers in Taiwan Mandarin and Chinese Mandarin.

References

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423. doi: 10.1002/j.1538-7305

Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1* (long papers) (pp. 1997–2005). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1181/ doi: 10.18653/v1/N18-1181