Indexing political identity through syntactic variables: Classifier specificity in pro-Taiwan and pro-China Taiwan Mandarin speakers

INTRODUCTION Political indexicality has been found in linguistic variables (cf. Hall-Lew & van Eyndhoven, 2025). Previous studies, however, focus on conspicuous phonetic and lexical variables. The present study examines whether a nuanced syntactic distinction — classifier specificity (whether a speaker uses a more specific/general classifier; henceforth CS; cf. 1a–1b) — may also serve to index the political identity among pro-Taiwan vs. pro-China Taiwan Mandarin speakers. Importantly, Hall-Lew & van Eyndhoven propose that political indexicality is typically derived from *politicizing* an existing indexicality. Intuitively, variety (Taiwan Mandarin vs. Chinese Mandarin) is the most likely target of such politicization. This suggests that if political identity is indeed indexed through CS, variety is also likely so.
METHODS Two questions were investigated through corpus analysis: 1) Is political identity indexed through CS in Taiwan Mandarin? and 2) If so, is the indexicality of variety a plausible target of such politicization? Corpus Corpora were built from 2,841 YouTube videos created by Taiwanese and Chinese content creators/news media. Stanford CoreNLP (Manning et al., 2014) was used to extract classifier-noun pairs. Classifier specificity CS was calculated as the mean absolute pointwise mutual information (PMI) of a classifier with all the nouns that it was found to occur with in the corpus (cf. Fig. 1). Social factor labeling Political identity was labeled based on the content creators/news media. Gender and age were included as control variables. A Wav2Vec2-based speech recognition model was first trained and then used to predict the speaker's gender/age. Statistical analysis Linear mixed-effects regression was used. In addition, two types of classifiers were identified through elbow analysis: general classifiers (the two most general classifiers, *ge* and *jian*) and specific classifiers (others). This was thus also included as a control variable.
RESULTS Female and older speakers were found to use more specific classifiers. Specifically, political identity was found to be correlated with CS among the Taiwan Mandarin speakers: pro-China speakers used more specific classifiers than pro-Taiwan speakers. In addition, variety differences were found. Taiwan Mandarin speakers used more specific classifiers than Chinese Mandarin speakers.
DISCUSSIONS Conventional social factors (gender and age) were found to be indexed through CS. This finding highlights the ability of nuanced variables to serve as socially meaningful variables. Importantly, political indexicality was found: pro-China Taiwan Mandarin speakers used more specific classifiers than pro-Taiwan speakers. This indexicality, however, may not come from the variety difference, which is the most common precursor of political indexicality in previous studies. While variety differences were indeed found, they were of the opposite directionality. Chinese Mandarin speakers used less specific classifiers than Taiwan Mandarin speakers. Instead, this political indexicality might come directly from the latent association of the more specific variants being more standard and formal. Under this view, it is possible that there is a discrepancy between the actual China and a glorified image of China for pro-China speakers, resulting in the use of the variants that might be perceived as more standard/formal.

(1) Examples of more general vs. more specific classifiers in Mandarin

a. A general classifier *ge*:	*Yi*	*ge*	*zhuo.zi.*	vs.	*Yi*	*ge*	*bei.bao.*
	one	CL:GENERAL	table		one	CL:GENERAL	backpack
	"a table"				"a backpack"		
b. A more specific classifier *zhang*:	*Yi*	*zhang*	*zhuo.zi.*	vs.	*Yi*	**zhang*	*bei.bao.*
	one	CL:SURFACE	table		one	CL:SURFACE	backpack
	"a table"				Intended: "a backpack"		

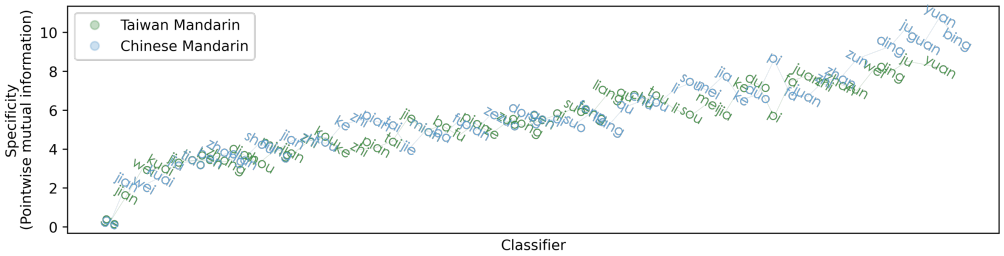

Figure 1. Calculated specificities of the classifiers in Taiwan Mandarin and Chinese Mandarin.

Table 1: Regression table for political identity (political identity: pro-China vs. pro-Taiwan; gender: female vs. male; age: younger vs. older).

Predictor	Est.	SE	df	*t*-value	*p*-value	
intercept	1.249	0.006	3902	226.043	<0.001	***
political identity	–0.055	0.008	1530	–6.81	<0.001	***
gender	–0.015	0.007	20530	–2.113	0.035	*
age	0.035	0.007	24570	5.073	<0.001	***
classifier type	3.239	0.008	26270	396.154	<0.001	***
political identity:gender	0.019	0.014	20430	1.345	0.179	
political identity:age	0.02	0.014	24600	1.5	0.134	
gender:age	–0.024	0.014	24900	–1.781	0.075	.
political identity:classifier type	–0.169	0.015	26360	–11.618	<0.001	***
gender:classifier type	–0.003	0.014	25910	–0.218	0.828	
age:classifier type	0.068	0.013	25710	5.068	<0.001	***
political identity:gender:age	–0.003	0.027	24880	–0.095	0.925	
political identity:gender:classifier type	0.01	0.027	25970	0.359	0.720	
political identity:age:classifier type	0.043	0.027	25690	1.61	0.107	
gender:age:classifier type	–0.026	0.027	25670	–0.971	0.331	
political identity:gender:age:classifier type	0.035	0.054	25660	0.655	0.512	

Table 2: Regression table for variety (variety: Chinese Mandarin vs. Taiwan Mandarin).

Predictor	Est.	SE	df	*t*-value	*p*-value	
intercept	1.053	0.003	19130	397.834	<0.001	***
variety	0.092	0.003	5366	29.849	<0.001	***
gender	–0.016	0.002	121400	–6.772	<0.001	***
age	0.039	0.002	187800	17.291	<0.001	***
classifier type	2.88	0.003	211300	1026.951	<0.001	***
variety:gender	0.0	0.005	121100	0.031	0.976	
variety:age	–0.029	0.005	187200	–6.399	<0.001	***
gender:age	0.008	0.004	195600	1.871	0.061	.
variety:classifier type	0.117	0.005	207300	24.561	<0.001	***
gender:classifier type	–0.012	0.004	208100	–2.657	0.008	**
age:classifier type	0.07	0.004	207800	15.809	<0.001	***
variety:gender:age	0.015	0.009	195400	1.727	0.084	.
variety:gender:classifier type	0.018	0.009	207900	2.029	0.042	*
variety:age:classifier type	–0.057	0.009	207700	–6.446	<0.001	***
gender:age:classifier type	0.039	0.009	207500	4.434	<0.001	***
variety:gender:age:classifier type	0.044	0.018	207600	2.493	0.013	*

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.

Hall-Lew, L., & van Eyndhoven, S. (2025). Linguistic variation and political identity. In C. Cieri et al. (Eds.), *Dimensions of linguistic variation* (online ed.). Oxford Academic.