The Extraction and Fine-grained Classification of Written Cantonese Materials through Linguistic Feature Detection

[1]Chaak-ming Lau, [2]Mingfei Lau, [1]Ann Wai Huen To
[1]The Education University of Hong Kong, [2]CanCLID
lchaakming@eduhk.hk, laubonghaudoi@icloud.com, towaihuenann@gmail.com

Abstract

This paper presents a linguistically-informed, non-machine-learning tool for classifying Written Cantonese, Standard Written Chinese, and the intermediate varieties used by Cantonese-speaking users from Hong Kong, which are often grouped into a single "Traditional Chinese" label. Our approach addresses the lack of textual materials for Cantonese NLP, a consequence of a lower sociolinguistic status of Written Cantonese and the interchangeable use of these varieties by users without sufficient language labeling. The tool utilizes key lexical markers identified from past linguistic research to determine whether a segment is Cantonese, Standard Written Chinese, mixed or unmarked. The task is reduced into string operations to allow for a flexible and efficient extraction of high-quality Cantonese data from large datasets mixed with Standard Written Chinese. This implementation ensures that the tool can process large amounts of data at a low cost by bypassing model-inferencing, which is particularly significant for marginalized languages. The tool also aims to provide a baseline measure for future classification systems, and the approach may be applicable to other low-resource regional or diglossic languages.

Keywords: Language Classifier, Cantonese, Diglossia

1. Introduction

Cantonese, a regional language prevalent in Hong Kong and parts of southern China, presents unique challenges and opportunities for the advancement of minority language resource development. Despite being a vibrant language with over 7 million users in Hong Kong (Census and Statistics Department, 2022; Bacon-Shone et al., 2015) and at least 40 million in nearby regions (Qu, 2021), it is currently considered a low-resource language (Joshi et al., 2020), notwithstanding its significant user base and clear economic demand.

The progress of Cantonese NLP has been disproportionately impeded due to the lack of appropriate written materials, a situation tied to the region's complex linguistic landscape. Like many low-resource languages with robust speaker communities, researchers have access to speakers and spoken materials but transcribed, written or labeled resources remain scarce. This scarcity is intensified by the diglossic situation in Hong Kong (Leung and Li, 2020), where most publicly available texts are written in Standard Written Chinese rather than Cantonese, or occasionally a blend of both. This situation is further complicated by copyright restrictions and the ineffectiveness of tools designed for Standard Chinese in accurately processing Cantonese.

The increasing need to compile resources for pre-training language models and generating automatic speech recognition training data is evident. An earlier version of this tool was first used as an efficient auto-classifier to mine Cantonese content from the vast amount of web data which contains a low percentage of Cantonese content. This paper further develops this method into a robust strategy that is devised based on past linguistic research. This paper first discusses a linguistic analysis of the "writing modes" involved in this classification task (§2), provides a linguistically-motivated task description (§3), and then presents a two-level rule-based implementation (§4) and an evaluation (§5) of the current library.

2. Cantonese and SWC

2.1. Contrasting the two varieties

The two main varieties under question are Cantonese (BCP 47: ***yue***) and Hong Kong Standard Written Chinese (SWC, BCP 47: ***zh-hk***). Both varieties are typically written in the Traditional Han Script (繁體中文). The former is usually used in speech, but it does cross the line occasionally: there is a higher chance of seeing Cantonese in informal writing, whereas the use of SWC is dominant in formal occasions. This is a case of diglossia (Ferguson, 1959), which refers to the use of two distinct varieties with different social statuses ("H" versus "L") and used in different social settings. The Hong Kong variant of SWC, often considered the "H" variety, is generally compatible with Mandarin and comprehensible to Chinese speakers outside Hong Kong. Cantonese dominates spoken communication, but is considered to be the "L" variety here. Its written form is unintelligible to non-users.

2nd EURALI Workshop, pages 24–29

Despite the apparent similarity between SWC and Mandarin, the two are significantly different in the Hong Kong context, as the former inherits some Cantonese lexical items and occasionally does not conform to Mandarin usage. Putonghua, the Mandarin-based national language of China, is seldom used among Hong Kong locals (See Li 2017, Leung and Li 2020 and Lai 2013), and therefore SWC is sometimes written by users who have zero Mandarin knowledge. Hong Kong SWC is filled with Cantonese elements in writing, which are analyzed as deviations from the Mandarin standard by some scholars (Shi et al., 2014; Tin, 2020), and simply a different register (or version) of Cantonese by others (Bauer, 1988; Snow, 2004, 2008).

Here is an example showing the difference between the two, and why this is not just a Cantonese versus Mandarin classification problem (See Lau 2024 for a full discussion). These sentences are modified from widely circulated examples found in teacher training materials in Hong Kong, which serve to illustrate the multiple writing norms used in Hong Kong. SWC words not accepted in spoken Cantonese are underlined. Cantonese elements that are SWC-violating are enclosed in boxes. Other elements without any special formatting are shared between SWC and Cantonese. LSHK Jyutping romanization is added on top of the characters.

(1) SWC	taa1 他	wo4 和	dai6dai6 弟弟	zo6 坐	haau6ce1 校車	soeng5hok6 上學
(2a) Can1	keoi5 佢	tung4 同	sai3lou2 細佬	daap3 搭	haau6ce1 校車	faan1hok6 返學
(2b) Can2	keoi5 佢	tung4 同	dai4dai2 弟弟	co5 坐	haau6ce1 校車	faan1hok6 返學
(3) Mixed	keoi5 佢	wo4 和	dai4dai2 弟弟	co5 坐	haau6ce1 校車	faan1hok6 返學
"*He and his younger brother go to school by school bus.*"						
(4) Unmarked			dai4dai2 弟弟	co5 坐	haau6ce1 校車	
"*Younger brother takes the school bus.*"						

Table 1: The spectrum between Cantonese and SWC

The SWC sentence (1) represents the norm taught in schools, distinct from everyday speech in (2a), mainly in terms of word choice. Despite this, they are pronounced in Cantonese using a nearly identical set of grapheme-to-phoneme conversion rules, rendering the sentence comprehensible, albeit unnatural-sounding, in spoken Cantonese. Some words are shared between SWC and Cantonese, for example the word for 'school bus' is shared, and the SWC words 'brother' and 'sit' are also legitimate in Cantonese. Sentence (2a) can be adjusted to resemble SWC more closely, as shown in (2b), without undermining its validity as a well-formed Cantonese sentence. Texts that mix the two, as in sentence (3), also exist. This sentence is not accepted in speech, nor is it recognized as SWC. This type of blending, or translanguaging, is commonplace in some use cases, e.g. texting. Conversely, there are sentences that are acceptable in both SWC and Cantonese, as shown in (4). This is a short sentence that does not contain any marked feature that will violate the convention of either SWC or Cantonese, and therefore usable in both forms.

The example above highlights the similarity between writing norms in Hong Kong, indicating that the non-binary nature of the problem. It is feasible, and indeed prevalent, for sentences or fragments to possess multiple statuses. This necessitates a carefully defined set of labels to better encapsulate the classification task.

2.2. The two varieties in computational linguistic literature

The classification of CJK languages has been of interest to the community. Work includes Xu et al. (2017); Huang and Lee (2008); Lu et al. (2020). However, most classification attempts focused on the major varieties and usually not the finer-grained distinctions, which is most needed in a minoritized language context. Cantonese and SWC have also been discussed in the literature on machine translation (Wong and Tsai, 2022). Previous work often presupposed a clear demarcation between the two, and resulted in a conversion between extreme points on the spectrum.

The Cantonese and SWC distinction, as illustrated above, with varying levels of social acceptance, is not as straightforward.

The ***NLLB*** (No Language Left Behind) project (Costa-jussà et al., 2022) developed a classifier designed to classify closely related languages. While it significantly contributed to the detection of sub-Saharan varieties, it struggled to accurately distinguish between Yue Chinese (***yue***, which is taken as Cantonese here) and Hong Kong Chinese (***zh-hk***), with results falling at chance level (p.33, figure 9). Upon further examination, this issue stems from the underlying FLORES dataset, which incorrectly labeled all SWC data as *yue*.

FastLangID[1] is a tool built on the original fastText model, emphasizing accurate classification between Asian languages. It supports three Chinese locales: Simplified Chinese (***zh-hans***), Traditional Chinese (***zh-hant***), and Yue Chinese (***zh-yue***). There is also a separate code for Cantonese

[1] https://github.com/ffreemt/fast-langid

(***yue***). The results from this library do not match the expectations of the task.

From this brief review, it is clear that the classification between Cantonese and SWC requires further scrutiny. This issue extends to many other underrepresented varieties. A bottom-up approach captures the differences between existing datasets, but determining where to draw the line (during data collection or labeling) requires top-down judgments from linguistic literature. This will be discussed in the subsequent section.

3. Linguistically-motivated task definition

Due to the noted inadequacy of a bottom-up approach, this section reviews the linguistic literature to reach a more accessible definition for the labeling of these closely related varieties. The challenge lies in determining a meaningful way for distinguishing between Cantonese and SWC.

Criterion 1: Text Comprehensibility Shi et al. (2014) base their classification on text comprehensibility by native, monolingual Mandarin speakers, suggesting that a text containing 50% or more incomprehensible Cantonese elements qualifies as Cantonese writing (p.6). This is, however, a negative definition that relies on the linguistic intuition of an external group of users, not Hong Kong users. For the classification task, the definition of SWC should capture the localized idealization of what the standard is like by Cantonese speakers, with some tolerance of local words. On the other hand, Cantonese is characterized by its authenticity as judged by its users, not by the existence of words unique to Cantonese, but by not using words that sound odd (i.e. violate the requirements).

Criterion 2: Distribution of Cantonese Elements Snow (2004) offers an in-depth analysis of the distribution of Cantonese and Standard Chinese elements in broadly-defined Cantonese writings, distinguishing six sub-types of Cantonese text based on how Cantonese is inserted. His work notably identifies intermediate mixing patterns (Random mixing, Patterned mixing, SWC narration with Cantonese dialogues), which are distinct document types requiring classification.

This paper uses the latter criterion as the basis for the classification task.

3.1. Language Labels for the Task

The two major categories in this task, Cantonese and SWC, and other related, intermediate varieties, are defined linguistically below based on the division of labor observed by speakers from Hong Kong.

	Example words
Cantonese feature	[嘅嗰啲咗佢喺咁啱有啲哋畀... 唔 [係得會好識使洗駛...
Cantonese exclude	(關係 | 吱唔 | 咿唔 | ...
SWC feature	[這哪唄咱啥甭那是的...
SWC exclude	是 [否日次非但旦]| ... [目綠藍紅中] 的 | 的 [士確式] ...

Table 2: A subset of items used for classification.

Cantonese A text that conforms to Cantonese speech in a non-verbatim reading process. Following this requirement, the use of SWC-marked elements will be a violation. Text under this category can be used in conversation.

SWC The school-taught written Chinese form, which is similar to Mandarin in many aspects but is read out in Cantonese. The writing process can be described as a replacement of words in Cantonese speech to eliminate disallowed elements (Lau, 2024).

Mixed A piece of text that contains random use of Cantonese and SWC elements, characterized by violation of both Cantonese and SWC requirements. For longer texts, there are two finer-grained labels: *"CantoneseInSWC"* and *"MixedInSWC"*, which refer to patterned insertion of Cantonese or Cantonese/SWC mixed segments in dialogues or quotes, while keeping SWC as the main language for the narrative.

Unmarked A string that does not show any features that clearly violate either Cantonese or SWC requirements.

3.2. Classification Approach

The core of the classification is keyword or key-string based, which is a variant of the bag-of-word strategy, but with units larger than words. This is also similar to the strategies used in LIWC (Pennebaker), widely used in social sciences research. Here is an abridged list[2] with features of Cantonese that clearly violate SWC, and vice versa. These features can be expressed in terms of lexical violations, which can be understood as elements that must not appear in the idealized varieties.

The features listed above are all lexical items. There are grammatical elements that Cantonese allows whereas SWC bans, such as the classifier-noun structure in the subject position (e.g. 隻狗 "CL-dog", 個袋 "CL-bag"). Such detection requires sentential parsing and may not bring significant

[2]A full list can be found in the project's public repository.

gain in document classification accuracy, and was therefore not implemented.

A segment is considered markedly Cantonese if it contains some Cantonese features, and does not contain SWC elements that violate the norm for Cantonese. A document is Cantonese if its constituent segments are either markedly Cantonese or Unmarked.

4. Implementation

Our proposed method has been implemented in Python and made publically available[3].

By default, regardless of the length of the document, classification will be done to the incoming string and a 4-way classification will be returned. This can be used for a short segment (e.g. a couple of sentences), or a longer document.

In the implementation, we first defined a list of Cantonese and SWC features in regular expressions (exemplified in Table 2) and the following variables:

1. $canto$: (# of Cantonese_Feature – Cantonese_Exclude) / Total_Features.

2. swc: (# of SWC_Feature – SWC_Exclude) / Total_Features.

3. $tolerance$: Highest acceptable percentage for a Neutral sentence, defaults to 0.01

4. $presence$: The threshold indicating "significant presence" of a variety, defaults to 0.03

5. $prevalence$: The difference between the ratio of two varieties that shall be counted as an overwhelming presence, defaults to 0.9

For each input segment, the number of Cantonese and SWC features are obtained by regex matches and classified into four classes based on the logic below:

For more accurate classification, two additional parameters can be set.

1. seg This option delimits all lines with clear punctuation marks (full stops, question marks, etc.) to obtain individual sentences. With multiple sentences, we can determine the category of the document more accurately. If a main category (either Cantonese or SWC) plus Unmarked sentences accounts for 95% of all segments, this will be returned as the label. If there is no clear winner, it will be returned as a Mixed document.

[3]https://github.com/CanCLID/cantonesedetect

Algorithm 1 Logic for Segment Judgment

if $canto$ + swc = 0 AND swc < $tolerance$ AND $canto$ < $tolerance$ **then**
  Unmarked
else
  if ($canto$ - swc) > $prevalence$ AND swc < $presence$ **then**
    Cantonese
  else if (swc - $canto$) > $prevalence$ AND $canto$ < $presence$ **then**
    SWC
  else
    Mixed
  end if
end if

2. $quotes$ This option divides the document into two parts: all text enclosed in a pair of quotation marks (quotes) and other text surrounding the quotes (matrix), the two sets will be sent to the classifier separately. This mode is particularly useful for the sub-categorization of Mixed writing, which is often done in a patterned manner, such as the use of Cantonese dialogues in an otherwise SWC text.

5. Evaluation

We constructed a test dataset with 420 sentences collected from published materials and social media from Hong Kong.

Table 3 shows some examples of this dataset. We first calculated the 4-way classification accuracy of our classifier, then we defined Cantonese as the positive label, thus the correct detection of Cantonese sentences as True Positives, and then calculated the confusion matrix and get the Precision and Recall results. Our experiments show that the 4-way classification accuracy can consistently remain 90%+.

5.1. Effectiveness of the tool

As mentioned above, the classifier in our experiments is implemented in a balanced way so that it doesn't put emphasis on any one of the 4 classes. However, since the original design goal was to extract Cantonese data from a large base of Chinese texts, we value its precision over recall, i.e. we prefer missing Cantonese sentences to misclassifying non-Cantonese sentences as Cantonese. Results of our evaluation are shown in Table 4. For other use cases where recall or overall accuracy is emphasized, one can adjust the classifier by adding/deleting the hard-coded linguistic feature list. For example, some elements like 和 ("and", as opposed to 同 in Cantonese) can be added as

Label	Number	Sentence examples
SWC	181	但這不應成為通車的阻礙 推廣心理和精神健康的重要性
Cantonese	59	就可以換購泰國直送嘅百分之百鮮芒果雪條 幫你輕鬆搵出全港最抵嘅貸款，甚至免息買二手車
Mixed	4	但長遠來講，都係申請息口較低的貸款比較划算 選定了心儀嘅機構先查詢個人實際年利率，咁會比較明智
Unmarked	176	如果你選擇租貸，就要預繳幾期供款 最低實際年利率：百分之五點一九
Total	**420**	

Table 3: Example sentences of our test dataset

a SWC feature for more aggressive filtering.

		Prediction	
		Cantonese	Non-Cantonese
Label	Cantonese	57	2
	Non-Canto	1	360
Precision			**0.983**
Recall			0.966
4-class accuracy			**0.967**

Table 4: Results of our approach on the test set

Our approach proved significantly better than existing methods, and is the first solution to effectively extract large-scale written Cantonese data for Large Language Model (LLM) and other downstream applications. Our approach reached 98.3%+ precision on our test dataset, which guarantees the extraction outputs are predominantly Cantonese.

On an AMD Ryzen 7 5800H CPU, our current implementation took 0.10 seconds to finish the classification of 420 sentences, compared to fastText's 0.48s with the `lid.176.bin` model.

Note that the current implementation is not fully optimized, and can be done so by implementing the strategy used in fastText.

5.2. Limitations

We acknowledge certain constraints of our language classification tool, listed as follows:

- **Precision:** The current implementation does not consider grammatical constructions, collocation and frequency. Adding more violation rules will give a higher precision.
- **Recall:** The tool may reject valid Cantonese or SWC expressions due to the use of certain strings in proper names that are not enclosed in quotes.
- **Workflow:** Codepoint-based filtering can be applied before determining the finer-grained distinctions.
- **Other varieties:** Currently the tool only classifies different genres used in Hong Kong, and does not take into account other forms of written Chinese varieties.

Despite these limitations, our tool demonstrates reasonable accuracy for the task. For our original use case which operates at the document level, multiple sentences form the basis of judgment. This ensures a fairly reliable classification. For a more purpose-general classifier, additional strategies can be added to further improve the tool's accuracy. This will be left for future work.

6. Conclusion

This paper discusses a classifier for Cantonese, primarily aimed at extracting relevant materials for training and beyond. While more sophisticated statistical or machine learning-based approaches could be employed, our rule-based approach utilizing simple string matching, has proven to be simple and high-performing.

A key insight of this solution is to approach language classification from research findings on vernacular writing, making a clear definition of language varieties. It is hoped that linguistically-motivated approaches will be considered in future task definitions for the classification of written forms of under-resourced languages.

7. Acknowledgments

We would like to express our sincere gratitude to the two anonymous reviewers for their valuable feedback, and members of the TypeDuck team at EdUHK for their continuous support.

8. Bibliographical References

John Bacon-Shone, Kingsley Bolton, and Kang Kwong Luke. 2015. *Language use, proficiency and attitudes in Hong Kong*. Social Sciences Research Centre, the University of Hong Kong, Hong Kong.

Robert S. Bauer. 1988. Written cantonese of hong kong. *Cahiers de Linguistique Asie Orientale*, 17(2):245 – 293.

Census and HKSAR Statistics Department. 2022. Main tables, 2021 population census.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 404–410, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mee Ling Lai. 2013. The linguistics landscape of Hong Kong after the change of sovereignty. *International Journal of Multilingualism*, 10(3):251–272.

Chaak Ming Lau. 2024. Ideologically driven divergence in cantonese vernacular writing practices. In Jean-François Dupré, editor, *The Politics of Language in Hong Kong*. Routledge.

Wai Mun Leung and David Chor Shing Li. 2020. 兩文三語: 香港語文教育政策研究 *[Biliteracy and Trilingualism: Language Education Policy Research in Hong Kong]*. City University of Hong Kong Press.

David Chor Shing Li. 2017. *Challenges in Acquiring Standard Written Chinese and Putonghua*, pages 71–107. Springer International Publishing, Cham.

Xugang Lu, Peng Shen, Yu Tsao, and Hisashi Kawai. 2020. Unsupervised neural adaptation model based on optimal transport for spoken language identification. *CoRR*, abs/2012.13152.

James W Pennebaker. Linguistic inquiry and word count: Liwc 2001.

Shao Bing Qu. 2021. 粵港澳大灣區語言生活狀況報告（*2021*）*[Report on the Status of Language Life in the Guangdong-Hong Kong-Macao Greater Bay Area (2021)]*. The Commercial Press.

Dingxu Shi, Jingmin Shao, and Zhiyu Zhu. 2014. 港式中文與標準中文的比較（第二版）*[Hong Kong Written Chinese and Standard Chinese: A comparison] (2nd ed.)*. Hong Kong Educational Publishing Co.

Don Snow. 2004. *Cantonese as written language the growth of a written Chinese vernacular*. Hong Kong University Press.

Don Snow. 2008. Cantonese as written standard? *Journal of Asian Pacific Communication*, 18(2):190–208.

Siu-lam [田小琳] Tin. 2020. 香港語言文字面面觀 *[Aspects of the language use in Hong Kong]*. Joint Publishing HK.

Ka Ming Wong and Richard Tzong-Han Tsai. 2022. Mixed embedding of xlm for unsupervised cantonese-chinese neural machine translation (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):13081–13082.

Fan Xu, Mingwen Wang, and Maoxi Li. 2017. Sentence-level dialects identification in the greater china region. *CoRR*, abs/1701.01908.