

## Grammatical Cues Enhance Accuracy of Word Segmentation: Evidence from Taiwan Hakka

This paper demonstrates that grammatical cues are crucial for accurate word segmentation in Taiwan Hakka. Using the multifunctional morpheme *mo* (無) as a case study, we show that systems relying only on dictionaries and manual corrections struggle with syntactic ambiguity. Although the segmentation system developed by the NCCU team reports an accuracy rate above 86%, our evaluation indicates that this performance is largely due to the extensive lexicon in the dictionary, rather than the system’s ability to analyze multifunctional words. For example, *mo* in Hakka can function as a negator (as in (1)), a question marker (as in (2)), or a disjunction marker (as in (3)).

- (1) Amin **mo** siid jiu.  
Amin not.have eat wine  
‘Amin does not drink wine.’
- (2) Amin siid jiu **mo**?  
Amin eat wine Q  
‘Does Amin drink wine?’
- (3) **Mo** Amin hi ho le.  
or Amin go good PRF  
‘Otherwise, Amin will go then.’

The NCCU segmentation system assigns the same tag, “VS (State Verb),” to all occurrences of *mo*, regardless of their syntactic environment. By contrast, our proposed grammar-sensitive segmentation system correctly distinguishes these functions by incorporating syntactic cues directly into the segmentation process. The outputs produced by our system accurately reflect the distinct grammatical roles of *mo* (FUNC\_inter marks a clause linker). See Table 1.

Table 1: Tagging results of *mo* in the two systems

	NCCU word segmenter	Grammar-sensitive word segmenter
<i>mo</i> in (1)	VS	FUNC negation
<i>mo</i> in (2)	VS	CLAUSE Q
<i>mo</i> in (3)	VS	FUNC inter

Remarkably, this level of accuracy does not require a large lexical base. Whereas the existing system relies on 21,617 entries from the MOE *Dictionary of Taiwan Hakka*, our system achieves high precision using only 6991 entries drawn from the Hakka Proficiency Test vocabulary lists. The key lies not in lexical quantity but in the implementation of grammatical rules that determine the function of *mo* according to its surrounding syntactic environment. Examples of such rules applied in our system include the following.

- Rule 1: *mo* → negation when followed by a verb, modifier, degree head, or modal
- Rule 2: *mo* → negation when followed by wh-elements
- Rule 3: *mo* → question marker when occurring in sentence-final position

These rules operate on the linearized outputs of the segmentation system, capturing grammatical distinctions without full syntactic trees or large annotated corpora. This approach aligns with linguistic theory and cognitive principles, as human comprehension relies heavily on grammatical information. Our results show that incorporating grammatical cues significantly improves segmentation accuracy for Taiwan Hakka, especially for multifunctional words, and suggest broader implications for low-resource languages, emphasizing grammar-sensitive over purely lexicon-driven methods.

*Selected references:* Hakka Affairs Council. Taiwan Hakka Corpus. Retrieved from <https://corpus.hakka.gov.tw/> Yeh, Chiou-shing, Huei-ling Lai, and Jyi-Shane Liu. 2021. The Construction of Taiwan Hakka Corpus and Preliminary Analysis of Hakka Lexical Usage, *Journal of Digital Archives and Digital Humanities* 8: 75-131.