

Uniform information density and the marking of Mandarin concessive-adversative relations: a surprisal-based study

Hongling Xiao, Sebastian Loftus, Marie-Catherine de Marneffe, Liesbeth Degand

Near-semantically equivalent alternations have been widely discussed from the perspective of uniform information density (UID) across multiple levels of linguistic organization (e.g. Levy & Jaeger, 2006; Jaeger, 2010; Asr & Demberg, 2015; Zhan & Levy, 2018). Under a UID view, speakers and writers tend to distribute information more evenly across an utterance, thereby avoiding sharp local peaks in processing difficulty. Explicit marking of concessive-adversative coherence relations in Mandarin Chinese provides a particularly suitable case, as the relation may be expressed either by a single conjunction in the first segment (e.g., 虽然 *suiran* ‘although’ P, Q) or by two conjunctions across both segments (e.g., 虽然 *suiran* ‘although’ P, 但(是) *dan(shi)* ‘but’ Q). Examples (1) and (2) illustrate typical patterns of single- and double-marking, respectively.

- (1) 虽然 [年龄相差一岁,]P [弟弟骡驹比哥哥马驹不仅显不出低矮, 而且比哥哥还要粗壮浑实。]Q
Although [their ages differ by one year,]P [the younger brother, a mule colt, is not only not shorter than his elder brother, a horse colt, but is also more robust and sturdier.]Q
- (2) 虽然 [她还没有完全公开地暴露自己的面目,]P 但是[学生们这样公开一闹, 毫无疑问校长会猜到有她在领导。]Q
Although [she hasn’t fully revealed her identity yet,]P **but** [with the students making such a public scene, the principal would undoubtedly guess that she was in charge.]Q

This study asks whether the choice between single- and double-conjunction marking is partly driven by the need to smooth information density at the semantic transition from P (the concessive) into Q (the adversative). Our primary hypothesis is that double marking is more likely when otherwise the onset of Q would be more surprising. We also consider two additional explanations. First, a multi-cue account predicts that the presence of concessive/adversative cues in Q reduces its reliance on a conjunction (see, e.g. Ding, 2008; Yin, 2020). Second, a complexity-based account predicts that greater complexity in P or Q segment—approximated here by token counts—favors explicit marking of the corresponding segment (Xing, 1992).

To test these hypotheses, we draw on naturally occurring instances of *suiran* constructions in fictional discourse from the CCL Corpus (Zhan et al., 2003), including both constructions with and without a Q-initial conjunction. The final dataset consists of 2,717 instances: 781 single-marking cases and 1,936 double-marking cases. Q-onset surprisal (in the absence of a Q-initial conjunction) was estimated using a Qwen3-1.7B-based language model (Qwen Team, 2025) under an 80/20 train-test split, with evaluation conducted on the test portion. We then fitted a logistic regression model with double marking as the outcome variable and Q-onset surprisal as the main predictor, controlling for P/Q segment length and the presence of Q-internal concessive/adversative adverbs.

Results show that three of the four predictors significantly contribute to the choice between single- and double-conjunction marking. The Q-onset surprisal trends in a direction consistent with the UID hypothesis: higher surprisal values indicate slightly higher odds of a second conjunction. This effect, however, did not reach statistical significance (OR=1.02, $p = .066$). Similarly, both P- and Q-segment length positively predict double marking, with longer segments reliably associated with higher odds of a second conjunction (respectively, OR=1.11, $p < .001$, and OR=1.08, $p < .001$). The strongest effect comes from the presence of additional concessive/adversative cues in Q, which sharply reduces the likelihood of double marking (OR=0.12, $p < .001$). Overall, these results do not provide strong evidence that double marking is directly driven by local Q-onset surprisal. Rather, the alternation appears to be shaped primarily by structural complexity and cue distribution.

Keywords: uniform information density; surprisal; double-conjunction marking

References

- Asr, F. T., & Demberg, V. (2015). Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th international conference on computational semantics* (pp. 118-128).
- Ding, Z. (2008). *Hanyu youbiao zhuanzhe fujie de guanlian biaoji moshi ji shiyong qingkuang kaocha* [An examination of the relational marking patterns and usage of marked adverbial complex sentences in Chinese] [Doctoral dissertation, Hunan Normal University].
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62.
- Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems* (Vol. 19).
- Qwen Team. (2025). *Qwen3-1.7B* [Large language model]. Hugging Face. <https://huggingface.co/Qwen/Qwen3-1.7B>
- Xing, F. (1992). Xiandai Hanyu zhuanzhe jushi [Modern Chinese adversative complex sentence patterns]. *Shijie Hanyu Jiaoxue [Chinese Teaching in the World]*, (2), 81–90.
- Yin, H. (2020). Xiandai Hanyu zhuanzhe guanxi xinlun [On adversative complex sentences in modern Chinese]. *Hanyu Xuebao [Chinese Linguistics]*, (1), 71–79.
- Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *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).
- Zhan, W., Guo, R., & Chen, Y. (2003). *The CCL corpus of Chinese texts*. http://ccl.pku.edu.cn:8080/ccl_corpus