Adjective ordering in Arabic: An analysis of the Wikipedia corpus

When more than one adjective describes a noun, these adjectives are usually used in a specific order. For example, a native English speaker is more likely to describe a table that is wooden and big as the "big wooden table" and not the "wooden big table". This phenomenon is called adjective ordering preferences, where certain adjectives are preferred closer to the modified noun than other adjectives. Adjective ordering preferences have been observed cross-linguistically in languages with pre-nominal adjectives (Dixon, 1982; Hetzron, 1978; LaPolla and Huang, 2004; Sproat and Shih, 1991), and languages with post-nominal adjectives (e.g., Indonesian; Martin, 1969). As to what predicts these preferences, Scontras et al. (2017) showed that these preferences are predicted by subjectivity, such that more subjective adjectives are preferred farther away from the noun. In the example "big wooden table", "big" here is more subjective than "wooden", so it occurs farther away from the noun. Subsequent work has further supported this proposal (Hahn et al., 2018; Samonte and Scontras, 2019; Kachakeche and Scontras, 2020). Recent behavioral work by Kachakeche and Scontras (2020) showed that native speakers of Arabic, a language with post-nominal adjectives, demonstrate stable adjective ordering preferences, and that these preferences are also predicted by subjectivity. The authors performed two behavioral experiments, one measuring adjectives' preferred distance from the noun, and another measuring adjective subjectivity amongst native speakers through a faultless disagreement task. In the current paper, we perform a corpus analysis to validate these behavioral results.

The aim of this paper is to understand how adjectives are ordered in naturalistic Arabic productions, and whether they match the behavioral results from Kachakeche and Scontras (2020). We provide a large-scale corpus study that establishes strong evidence for subjectivity-based adjective ordering in Arabic productions. The analysis is conducted on the Wikipedia corpus, a large dataset of encyclopedia articles, open sourced by Wikipedia. Our work takes advantage of several tools for data collection and analysis. The WikiExtractor (Attardi, 2012) was used to extract plain text from the raw Wikipedia data, and the Farasa POS tagger (Abdelali, Darwish, Durrani, and Mubarak, Abdelali et al.) was used to automate part-of-speech tagging. The Farasa tagger was created by Qatar Computing Research Institute, and is designed specifically for Arabic; an initial attempt using the Stanford tagger (Toutanova et al., 2003) took much longer and led to lower accuracy. We extracted all adjective-adjective-noun phrases from the Wikipedia corpus (1,892,619 triples). From these phrases, we considered the ones that contained at least one of the adjectives used in the experiment by Kachakeche and Scontras (2020). These selection criteria led to 91,564 adjective-adjective-noun triples. For each of the adjectives of interest, we computed a single average-distance measure, which indicates how far away the adjective is from the modified noun. Figure 1 plots these distance scores against the scores from the behavioral measure by Kachakeche and Scontras (2020), which account for 69% of the variance in the corpus measure. Given this strong correlation, we can conclude that the behavioral judgment data collected by Kachakeche and Scontras (2020) tap into the same preferences that speakers use as they produce multi-adjective phrases. Figure 2 plots the results from the corpus analysis against the faultless disagreement subjectivity scores, demonstrating that adjective subjectivity predicts Arabic ordering preferences also in naturalistic productions.

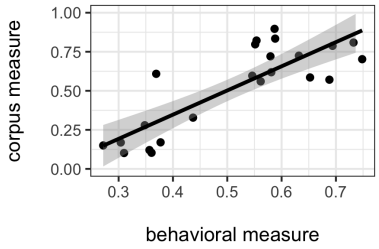

Figure 1. Results from the corpus analysis plotted against behavioral results by Kachakeche and Scontras (2020) for each of the 25 adjectives tested. Corpus results strongly correlate with the behavioral results ($r^2 = 0.69$, 95% CI $[0.42, 0.83]$).

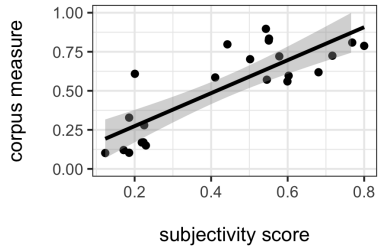

Figure 2. Results from the corpus analysis plotted against faultless disagreement scores from Kachakeche and Scontras (2020). Corpus results strongly correlate with the faultless disagreement scores ($r^2 = 0.68$, 95% CI $[0.42, 0.82]$).

Abdelali, A., K. Darwish, N. Durrani, and H. Mubarak. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations*.

Attardi, G. (2012). Wikiextractor. `https://https://github.com/attardi/wikiextractor`.

Dixon, R. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Hahn, M., J. Degen, N. D. Goodman, D. Jurafsky, and R. Futrell (2018). An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th annual conference of the Cognitive Science Society*, London. Cognitive Science Society.

Hetzron, R. (1978). On the relative order of adjectives. In H. Seiler (Ed.), *Language Universals*, pp. 165–184. Tübingen, Germany: Narr.

Kachakeche, Z. and G. Scontras (2020). Adjective ordering in Arabic: Post-nominal structure and subjectivity-based preferences. In *Proceedings of the Linguistic Society of America*, Volume 5, pp. 419–430.

LaPolla, R. J. and C. Huang (2004). Adjectives in Qiang. In R. M. Dixon and A. Y. Aikhenvald (Eds.), *Adjective Classes: A Cross-Linguistic Typology*, pp. 306–322. Oxford: Oxford University Press Oxford.

Martin, J. E. (1969). Some competence-process relationships in noun phrases with prenominal and postnominal adjectives. *Journal of Verbal Learning and Verbal Behavior 8*(4), 471–480.

Samonte, S. and G. Scontras (2019). Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. In *Proceedings of the Linguistic Society of America*, Volume 4, pp. 1–13.

Scontras, G., J. Degen, and N. D. Goodman (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science 1*(1), 53–65.

Sproat, R. and C. Shih (1991). The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos and R. Ishihara (Eds.), *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, pp. 565–593. Kluwer Academic Publishers.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology-volume 1*, pp. 173–180. Association for Computational Linguistics.