Revisiting the idiom principle through the lens of an agglutinating language: A corpus-based description of adjective-noun collocations in Turkish and English

Doğuş Can Öksüz and Vaclav Brezina (Lancaster University, UK)

Most research into formulaic language has been limited to a narrow set of languages particularly English (Durrant, 2013). For example, Sinclair's (1991) "idiom principle", which states that a language user has large number of available semi-preconstructed phrases that constitute single choices during processing, - has only rarely been applied to highly inflected, agglutinative languages such as Turkish. For this reason, our knowledge on the role they play in formal, functional, pragmatics and processing aspects, has been restricted to the narrow set of languages and the status of formulaicity as a property of language as such has not sufficiently established (Durrant, 2013). Biber (2009) suggested that agglutinating languages such as Turkish and Finnish are particularly interesting candidates for investigating the scope and status of formulaicity as a property of language. In this paper, therefore, we demonstrate that the formulaicity in agglutinating languages such as Turkish is different from the formulaicity in non-agglutinating languages. One of the reasons is that the rich morphology of an agglutinating language such as Turkish, affects the frequency of occurrence and syntagmatic associations between lexical items.

This study, following a frequency-based tradition of collocation research (Evert, 2008, Hoey, 2005; Sinclair 2004), is a corpus-based description of two-word adjective-noun collocations in Turkish and English. Corpora chosen for this study represent the input that language users experience on a daily base. The Turkish National Corpus (TNC) is a written, and general corpus of Turkish with a size of 47,641,688 tokens. It is a collection of 4438 different text samples, representing 9 domains and 34 different genres, written in between 1990-2009. The British National Corpus (BNC) XML edition was used as a comparison corpus. To extract the adjective-noun collocations, frequency bands were established making use of the TNC and the BNC word frequency lists. For establishing the bands, the frequency scale of the nouns in the TNC and the BNC were considered and five node words were selected from high and mid-frequency bands. Only the nouns were selected as node words to search for adjective collocates. Using the raw frequency scores, the most frequent four adjective collocates of each node words, a total of forty collocations were extracted for further analyses on two-word collocations' frequency of occurrence and collocational strength. The adjective-noun collocations were chosen as focus of the investigation for two reasons: First, nouns within adjective-noun collocations can be inflected with various types of suffixes including case marking, plural and instrumental in Turkish, and thus it is possible to observe the influence of agglutination on the collocability of adjectives and nouns. Second, they occur in a certain syntactic order in which adjectives precede the nouns in both Turkish and English, as observed in the concordances lines; hence they should be

fully comparable for both strength and directions of the syntagmatic associations in Turkish and English.

For the frequency comparisons, raw frequency scores of adjective-noun collocations in the BNC and the TNC were relativized to per million words. This allows a comparison of how many times a two-word adjective-noun collocation in Turkish is likely to occur per million words against its equivalent in English. Besides the frequency comparison, this study focused on the collocational strength, the association between the node words and the collocates as measured by Mutual information (MI), Log Dice and Delta P metrics. The collocational relationship is quite complex and no single measure of association can capture the full complexity of this relationship (Brezina, McEnery, & Wattam, 2015). Therefore, this research used three different corpus derived measures of associations to investigate the collocational strength. MI scores were used to measure the rare exclusivity, that is, MI score tends to highlight the relatively infrequent words with low co-occurrence frequency (Evert, 2008, Manning & Schütze, 1999). Thus, another measure of association, which is neutral to the low frequency of occurrence, was needed. In this regard, Log Dice provides both an accurate measure of association and easily interpretable scale of scores (Rychly, 2008). MI and Log dice measures consider collocational strength as necessarily symmetrical. For this reason, Delta P was used as a measure of probability that takes directionality of the collocational strength into account. In this study, Delta P was used as a measure of association alongside MI and Log Dice measures to investigate whether adjectives are more predictive of the following nouns or the nouns are more predictive of the preceding adjectives (Gries, 2013).

Lemmatisation of the nominal inflections within adjective-noun collocations to abstract away from the complex morphology of Turkish was a natural step. For calculating the lemmatised forms of the collocations' relative collocate frequency and association scores in Turkish and English, the raw frequencies of the node words' all of the inflected forms and the collocations' all of the inflected forms were identified. The frequency sums of the inflected forms of the node words were taken as frequency of the node in the whole corpus and the frequency sums of the all inflected forms of the collocations were taken as frequency of the collocation in the collocation window. Thus, the relative collocate frequency and association scores, as measured by MI, Log Dice and Delta P, were calculated for both collocations' inflected and uninflected forms in Turkish English. After calculating the relative frequency and association scores for each inflected forms, the same measures were calculated for the lemmatised form of the collocations.

Overall, this study made possible to observe the relative collocate frequency and association scores of each inflected forms of the adjective-noun collocations. Furthermore, it was possible to make a comparison between the lemmatised and unlemmatised forms of the adjective-noun collocations for frequency of occurrence and associations in Turkish and in English. The data revealed that (55%) of high-frequency band, and (50%) of mid-frequency band collocations reach higher unlemmatised relative collocate frequency scores in English than their equivalents in Turkish. Lemmatised relative collocate frequency scores indicate that (70%) of high-frequency

band and (75%) of mid-frequency band collocations reach higher scores in Turkish than their equivalents in English. This may be viewed as a surprising finding considering the fact that lemmatised forms of the collocations in English and Turkish are functionally equivalent. Unlike relative collocate frequency scores, the adjective-noun collocations reach predominantly higher unlemmatised and lemmatised MI scores in English than in Turkish. Unlemmatised MI scores indicate that only (20%) of high-frequency band and (40%) of mid-frequency band adjective-noun collocations reach higher MI scores in Turkish than their equivalents in English. Similar to unlemmatised MI scores, only (25%) of high-frequency band, and (40%) of mid-frequency band adjective-noun collocations reach higher lemmatised MI scores in Turkish. To conclude, adjective-noun collocations in Turkish tend to be more frequent, but weaker associated than their equivalents in English. This might have implications that speakers of Turkish and English might need to process adjective-noun collocations differently in their respective L1s.

References

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Ufuk, U., ... Yıldız, I. (2012).Construction of the Turkish National Corpus (TNC). *Lrec 2012*, 3223–3227.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275– 311. http://doi.org/10.1075/ijcl.14.3.08bib

BNC: The British National Corpus. http://bncweb.lancs.ac.uk/

Brezina, V., Mcenery, T., & Wattam, S. (2015). Collocations in context A new perspective on collocation networks*. *International Journal of Corpus Linguistics*, *202*(2015), 139– 173. http://doi.org/10.1075/ijcl.20.2.01bre

Durrant, P. (2013). Formulaicity in an agglutinating language: the case of Turkish. *Corpus Linguistics and Linguistic Theory*, *9*(1), 1–38. http://doi.org/10.1515/cllt-2013-0009

Evert, S. (2008). Corpora and collocations. In A. Ludeling & Merja Kyto (Eds.), *Corpus Linguistics. An international handbook* (pp. 1212–1249). Berlin and New York: Mouton de Gruyter.

Gries, S. T. (2013). 50-something years of work on collocations What is or should be next ...*. *International Journal of Corpus Linguistics*, *1*, 137–165. http://doi.org/10.1075/ijcl.18.1.09gri

Hoey, M. (2005). *Lexical priming: A new theory of words and language* (1st ed.). London: Routledge.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing* (2nd ed.). London: MIT Press.

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Languages Processing* (pp. 6–9). Retrieved from http://nlp.fi.muni.cz/raslan/2008/raslan08.pdf#page=14

Sinclair, J. (1991). *Corpus condordance collocation* (1st ed.). Oxford: Oxford University

Press.
Sinclair, J. (2004). *Trust the text language, corpus and discourse*. (Ronald Carter & J. Sinclair, Eds.). London: Routledge.