Prompted by advances in corpus studies and sophistication in available software, current research revealed significant aspects of frequency in emergence and use of lexical patterns. It has been argued that operations of rules are defined by lexical constraints and the division between lexicon and grammar cannot be maintained (Bybee and Hopper, 2001; Sinclair, 1991, 2004; Sinclair and Mauranen, 2006) Since corpus analyses combine quantitative and qualitative methodologies, we now know that “4000-5000 most frequent words account for up to 95 percent of a written text and 1000 most frequent words account for 85 percent of speech.” (Rayson and Davies, 2010:vii) Furthermore, quantitative analyses concluded that more than half of the units in ordinary language are pre-fabricated (formulaic) expressions (Wray, 2008).

Previous research on frequency of lexical items concentrated on distributional properties, processing and storage. Current corpus-based frequency studies, however, showed that multi-word units are also frequent and that language users are sensitive to frequencies of multi-word, the frequent items are processed faster. It is observed that effect of frequency of such units cannot be reduced to frequency of individual items in a multi-word unit (Arnon and Snider, 2010).

In an earlier study, we have provided a structural typology of Turkish multi-word units and discussed their internal composition and pragmatic functions (Aksan and Aksan, 2013). In this study, we examine the role of frequency in patterning of multi-word units in Turkish. To achieve this goal, multi-words units emerging in 10-million words corpus constructed from written and spoken datasets of Turkish National Corpus (http://www.tnc.org.tr) is analyzed both automatically and manually. Ngram Statistical Package software tool (Banerjee and Pedersen, 2003) is used to generate rank order frequency lists of n-grams. The data show that among the most frequent bi-grams (ya da, ve bu, hem de, böyle bir, etc.) we rarely find any “content” words, as in the case of many other languages. Similarly, the most frequent tri-grams are also composed of function words (bir süre sonra, bir kez daha, ne var ki, her ne kadar, etc.) and serve more like formulaic expressions. The four-gram units (her zaman olduğu gibi, başka bir şey değildir, ve buna bağlı olarak, etc.) appear to include more content words than others, however they include tri-gram units, and are less frequent in the 10-million words corpus.

In brief, this study argues that studying frequency effect in patterning of multi-word units may shed light on characteristics of formulaicity in Turkish.

References
Banerjee, S., T. Pederson (2003). The design, implementation and use of the (N)gram (S)tatistic (P)ackage. Proceedings of the fourth international conference on intelligent text processing and computational linguistics, 370-381.

* This study is supported by TÜBİTAK (Grant no: 113K039).