Individual word and phrase frequency effects in collocational processing: Evidence from typologically different languages, English and Turkish

Doğuş Öksüz[1] Patrick Rebuschat[2], Vaclav Brezina[3]

[1] University of Cambridge,
[2]Lancaster University
Correspondence concerning this paper should be addressed to Doğuş Öksüz (dco24@cam.ac.uk)

Usage-based approaches to language learning view multi-word sequences (MWS) as essential building blocks for language learning and processing (e.g. Arnon, McCauley & Christiansen, 2017). MWS include collocations (e.g. *front door*), binomials (*bread and butter*), and idioms (*kick the bucket*). Importantly so far, the vast majority of psycholinguistic experiments have focused on a narrow range of primarily European languages, which makes it difficult to generalize the findings to other languages (Durrant, 2013). In this paper, we focus on the effect of linguistic typology on the processing of collocations – a prominent type of MWS. We conducted a corpus analysis alongside psycholinguistic experiments to examine the processing of adjective-noun collocations in Turkish and English by native-speakers of the two. Turkish is an agglutinating language with a rich morphology, building up complex word forms. This prompts questions about collocational processing in typologically different languages around the frequency effects of individual words and whole phrases: are collocations processed similarly across languages or do they require different processing depending on their typological characteristics?

Conducting a contrastive corpus study, we investigated the extent to which frequency counts and association statistics are different for Turkish and English adjective-noun collocations (e.g. *middle class*/*orta sınıf*). Using comparable, and balanced reference corpora of the two languages, the BNC for English and the TNC for Turkish, we firstly examined the differences in collocations' frequency counts and association statistics between lemmas and word forms. In addition, we assessed to what extent high-frequency inflected collocations exist in Turkish. Poisson regression modelling showed that base-form Turkish collocations have significantly lower frequency counts than English ones, because the base-form collocations in English potentially subsume the Turkish equivalents of both the base and its inflected forms. With regard to the lemmatized collocations, the vast majority occurred at a higher-frequency than their English equivalents. In addition, the agglutinating structure of Turkish appears to increase adjective-noun collocations' association statistics.

We conducted online acceptability judgment tasks to explore the sensitivity of native speakers of English (*n=30*) and Turkish (*n=46*) to the frequencies of adjectives, nouns and whole collocations. A total of 120 adjective-noun combinations were extracted each from the BNC and TNC: (1) high-frequency collocations (e.g. *dark hair*), (2) low-frequency collocations (*lovely house*), and (3) baseline items (*general eyes*). Mixed-effects regression modelling revealed that speakers of both languages processed adjective-noun collocations at similar speeds (see Figure 1). Alongside collocation frequencies English native-speakers were sensitive to noun frequencies, which led to faster response times. However, lemmatized frequencies of nouns led to slower response times for Turkish speakers. Also, Turkish speakers were not sensitive to the non-lemmatised noun frequency counts. Taken together, the evidence suggest that speakers of both languages are found to be chunking individual words into MWS - they were sensitive to the phrasal frequency information, frequently co-occurring adjacent elements are easily chunked, facilitating processing (Christiansen & Chater, 2016). However, collocational processing also depends on language-specific usage-based constraints that vary cross-linguistically. Processing MWS can be described as a probabilistic graded phenomenon that is affected by language-specific factors.

References

Arnon, I., McCauley, S.M. & Christiansen, M.H (2017). Digging up the building blocks of language: Age-of-acquisiton effects for multiword phrases. *Journal of Memory and Language* 92, 12(1): e0168532

Chater, N. & Christiansen, M.H. (2016). Sequeezing through the Now-or-Never bottleneck:Reconnecting language processing, acquisition, change and structure. *Behavioral & Brain Sciences, 39,* e62

Durrant, P. (2013). Formulaicity in an agglutinating language: The case of Turkish. *Corpus Linguistics and Linguistic Theory, 9*, 1–38.

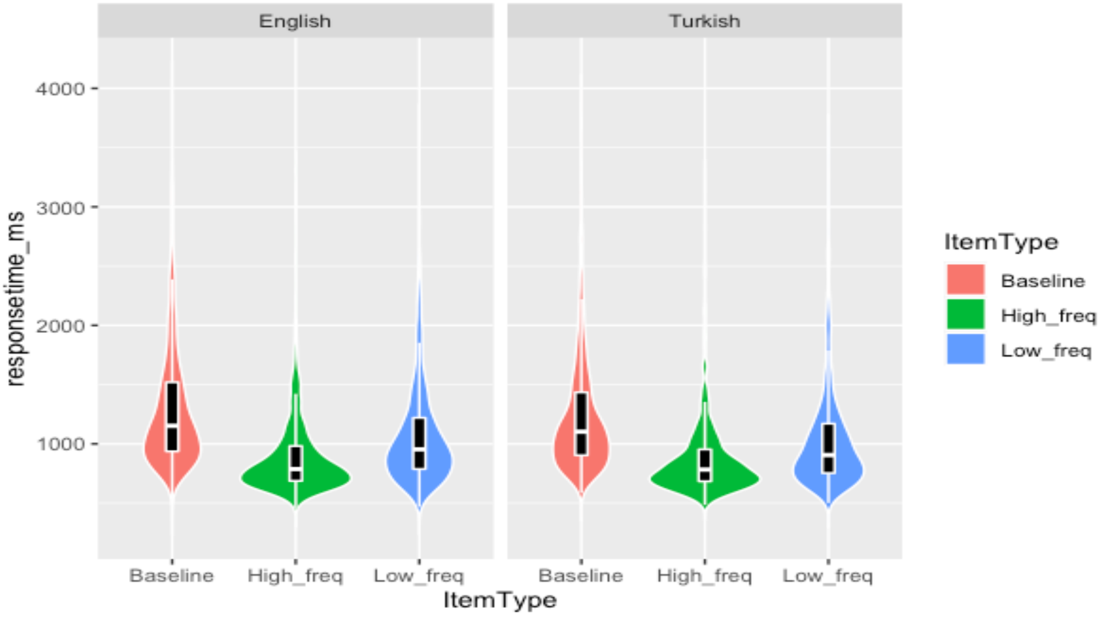

Figure 1. Distribution of response times for item types in English and Turkish