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

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, especially English, which makes it difficult to generalize the findings to other languages. In this paper, we focus on the effect of linguistic typology on the processing of collocations – a prominent type of MWS due to their frequency and versatility. The paper conducts 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, building up complex word forms, utilizing remarkably productive morphology. 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. 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. This shed light on how the complex morphology of Turkish affects collocational relationships. 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 because lemmatized forms are more strongly associated than their base forms.

We conducted online acceptability judgment tasks to explore how English (*n=30*) and Turkish (*n=46*) native-speakers process adjective-noun collocations. We specifically focused on whether speakers of English and Turkish (1) process adjective-noun collocations with comparable speeds in their respective native languages, (2) are sensitive to single word and phrasal frequency information simultaneously, and (3) are sensitive to frequency information differs depending on the frequency of the collocations.

A total of 120 adjective-noun combinations were extracted each from the BNC and TNC. The items fell into one of three critical conditions: (1) high-frequency collocations (e.g. *dark hair*), (2) low-frequency collocations (*lovely house*), and (3) non-collocational (baseline) items (*general eyes*). Individual word frequency, collocation frequency counts, and association statistics of the items were obtained from the two corpora. 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 adjective and 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. This provides support for the elevated status of MWS as a general feature of language; 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. Turkish collocations are found to be processed more holistically since Turkish speakers were less sensitive to the individual word-level frequency information than English speakers. 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

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioural & Brain Sciences, 39,* e62.

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

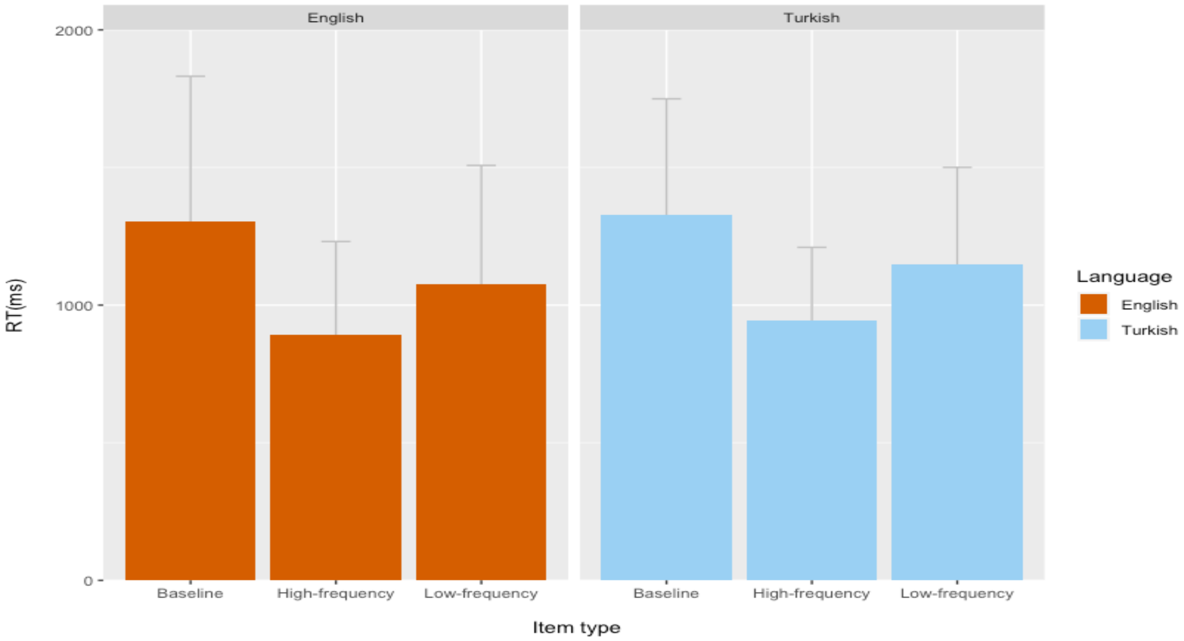

Figure 1. Response times for item types in English and Turkish