USAGE OF TAXONOMICALLY ENRICHED TURKISH LEARNER CORPUS

This chapter illustrates the practical use of the Taxonomically Enriched Turkish Learner Corpus, emphasizing the benefits its taxonomic structure and annotation extender offer to users. To demonstrate how individual facets help reveal the underlying nature and characteristics of learner errors, CASE errors, which constitute one of the most frequent error types and are associated with an important morphological property of Turkish, are examined as a representative example.

First, we can explore how CASE errors correlate with the POS facet to identify specific patterns in the data. Figure 1 presents the frequency distribution of POS tags associated with CASE errors. This distribution provides a clear overview of which linguistic categories are most susceptible to case-marking deviations. The vast majority of CASE errors occur within the nominal category, accounting for 71.2% of all instances (comprising 69.5% common nouns and 1.7% proper nouns). VERB (10.1%) and PRON (9.3%) follow as the next most significant contributors. The VERB category reflects nominalized verb forms, such as verbal nouns and participles. All other POS labels account for a substantially smaller proportion of the total errors.

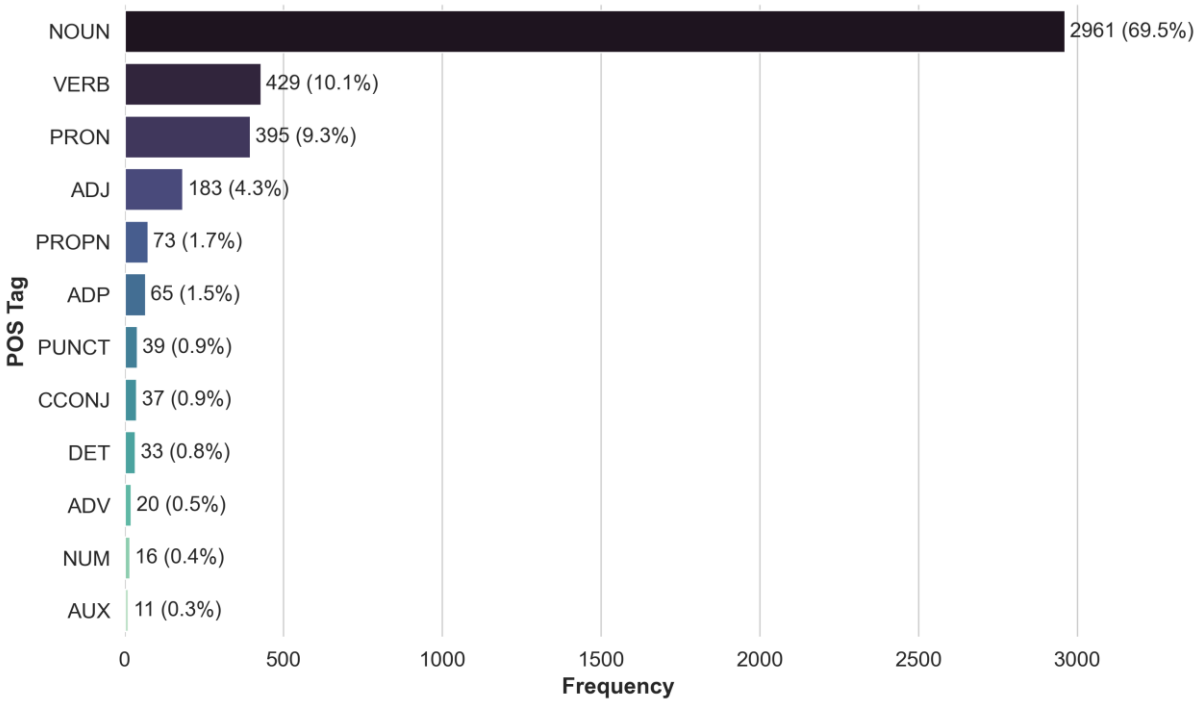

Figure 1 Distribution of POS subfacet for CASE Errors

Through a more detailed analysis, the distribution of CASE errors across the Phenomenon facet can be examined to determine which phenomenon (omission, addition, or misuse of case affixes) is more prevalent among learners. As shown in Figure 2, the results reveal a predominance of Omission, which accounts for 58.1% (N=2,361) of all CASE errors. Addition emerges as the second most frequent category (N=1,142; 28.1%), while Misuse constitutes the smallest portion of the dataset (N=561; 13.8%). The predominance of Omission errors (58.1%) can be largely attributed to the typological characteristics of Turkish as an agglutinative language with a zero-marked nominative case. In Turkish, the nominative case represents the base form of the noun and does not take any overt affix. Consequently, when learners face uncertainty regarding case assignment or cognitive overload during morphological processing, they tend to default to this unmarked bare form. This strategy results in the omission of required affixes.

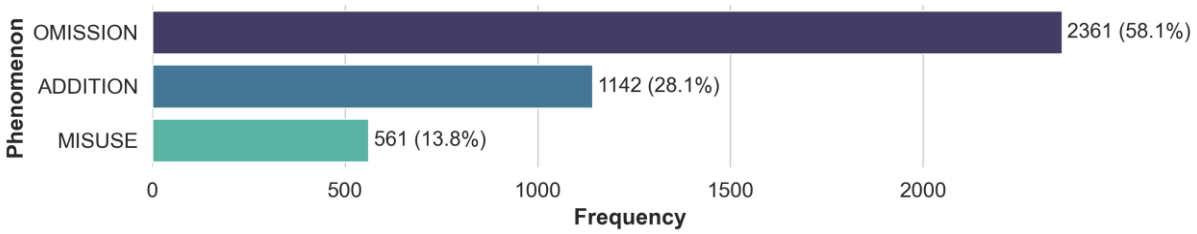

Figure 2 Distribution of Phenomenon facet for CASE Errors

To investigate how CASE errors interact with other grammatical categories, we can examine them in the context of the Inflectional Features facet. Figure 3 presents the distribution of inflectional feature labels associated with the tokens identified in CASE errors. The analysis reveals that the errors are overwhelmingly concentrated on nouns exhibiting the default, unmarked morphological values. In particular, Person=3 (N=3,561) and Number=Sing (N=2,604) constitute the most frequent features, accounting for 22.0% and 16.1% of the total tag count, respectively. Regarding specific case markers, the Accusative case (Case=Acc) is the most frequent explicitly realized case feature (N=1,165; 7.2%), followed by the Nominative (Case=Nom; 5.0%) and Genitive (Case=Gen; 4.6%). Notably, the data also indicates a significant prevalence of possession-related markers such as Number[psor]=Sing (6.8%) and Person[psor]=3 (6.3%), suggesting that CASE errors are frequently intertwined with the morphological complexity of possessive constructions. The remaining low-frequency features are grouped under the Others category, which comprises 3.4% (N=553) of the distribution.

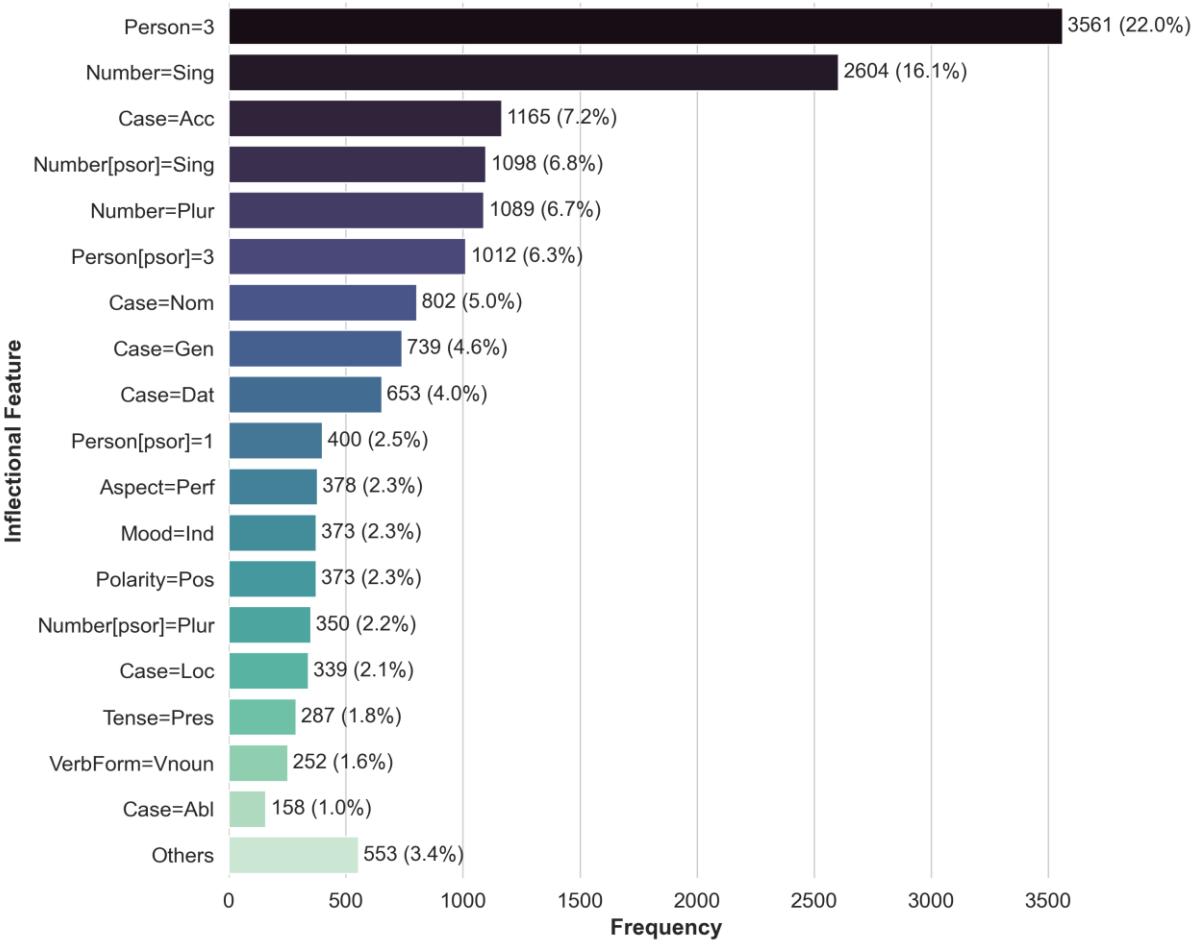

Figure 3 Distribution of Inflectional Feature facet for CASE Errors

As previously noted, 9.3% of CASE errors occur within the pronoun (PRON) category. Should researchers wish to investigate which specific types of pronouns are most prone to these errors, the Lexical Features facet can be employed to provide a more granular breakdown. Figure 4 illustrates that personal pronouns (PronType=Prs) emerge as the most frequent category, accounting for nearly half of all tagged lexical features (N=168; 48.0%). This is followed by reflexive pronouns (Reflex=Yes), which constitute 22.6% (N=79) of the instances, and demonstrative pronouns (PronType=Dem) at 20.3% (N=71).

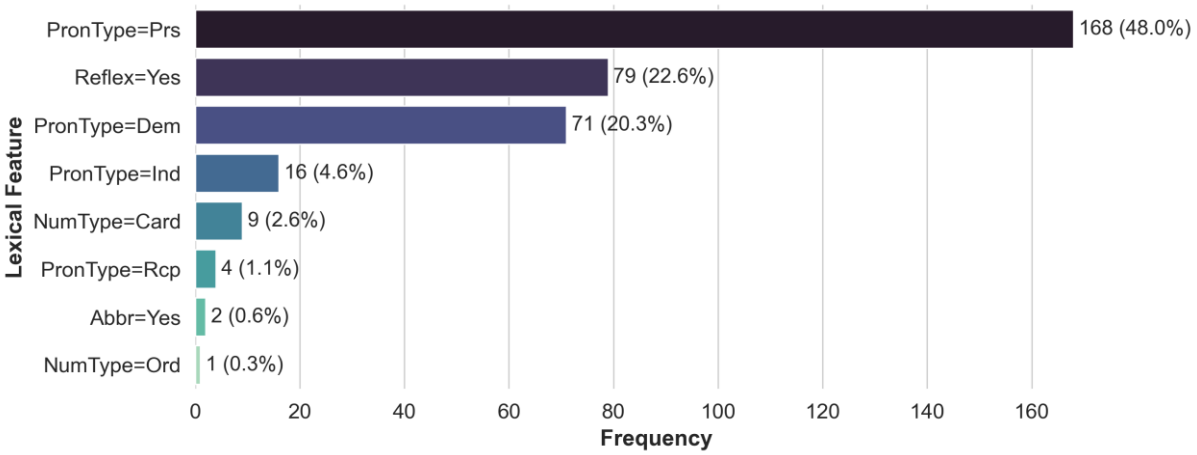

Figure 4 Distribution of Lexical Feature facet for CASE Errors

Understanding the L1 background of learners can be also significant, as it allows for the identification of language-specific patterns. This helps distinguish between errors resulting from L1 transfer and those that are language-independent and shared by learners from diverse linguistic origins. Figure 5 presents the distribution of the top 10 nationalities associated with CASE errors. The data indicate that learners from Syria constitute the largest group, accounting for 9.0% (N = 364) of the total error count, followed by learners from Turkmenistan (7.2%; N = 293) and Afghanistan (5.8%; N = 235). A closer examination of these linguistic backgrounds reveals three broad clusters: Arabic-speaking (Syria, Iraq, Morocco), Indo-Iranian (Afghanistan, Iran), and, notably, Turkic (Turkmenistan, Azerbaijan, Kazakhstan). While the prevalence of errors among Arabic and Persian speakers is expected due to profound typological differences, the high frequency of errors among speakers of cognate languages is particularly noteworthy. Specifically, the prominence of learners from Turkmenistan (#2) and Azerbaijan (#6) suggests that despite their typological proximity to Turkish, these learners may experience negative transfer arising from subtle but systematic morphological divergences between their native Turkic varieties and the case-marking conventions of standard Turkish.

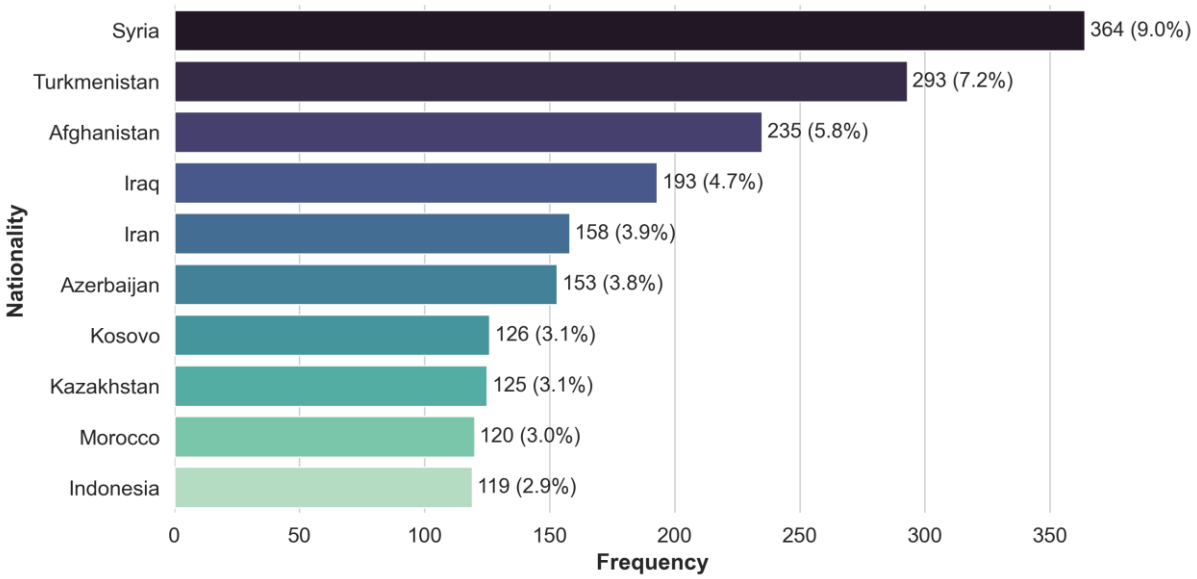

Figure 5 Distribution of Nationality for CASE Errors

It is crucial to interpret demographic data within a broader contextual framework. Finally, it should be noted that the prominence of Syrian learners in this distribution is likely influenced, at least in part, by their substantial demographic representation within the overall learner corpus. The integration of a user interface can leverage the multi-layered taxonomic framework to permit versatile search combinations. Such a system will make it possible to retrieve highly specific data points, for example, identifying the over-application of accusative marking in reflexive pronouns among Kosovar students by cross-referencing the Case, Phenomenon, Inflectional Features, Lexical Features, and Metadata facets.