Tokenization Bias in Corpus Tools and Its Consequences for Lexical Density Measurement

Lexical density, introduced by Halliday (1985) as the proportion of content words relative to the total number of clauses or words in a text, works reasonably well for English, for which it was developed. Applying it to Croatian exposes methodological problems that have not received sufficient attention.

This paper examines what happens when lexical density is calculated for Croatian using pipelines whose tokenization principles were developed around English. Parallel Croatian and English texts were processed with automatic tokenization and POS tagging (CLASSLA-Stanza, whose output is displayed in interfaces such as Sketch Engine) and then manually re-annotated to establish a corrected baseline; the scale and direction of the effects are illustrated on a 2.7-billion-token Croatian web corpus.

Croatian presents categorization problems that do not exist in English. The reflexive particles *se* and *si* are part of a reflexive verb in some contexts and mark the passive or an impersonal construction in others, and enclitic auxiliaries alternate with full lexical verbs (the enclitic *ću* in *Doći ću* 'I will come' versus the stressed *hoću* 'I want (it)'). Crucially, these elements are not misclassified: *se* is tagged as a pronoun in over 99% of its 37.9 million occurrences. Their separate tokenization adds grammatical tokens to the denominator and therefore lowers, rather than raises, the measured lexical density of Croatian texts. Negation works in the opposite direction. The particle *ne* is written separately from finite verbs in general, but the present-tense forms of *htjeti, biti* and *imati* are conventionally fused (*neću*, *nisam*, *nemam*), so roughly two thirds of negated verb forms are one token and one third two tokens for the same morphosyntactic configuration. The annotation compounds this: the tagset provides a Negative attribute for verbs but never uses it, so *nije* and *je* receive identical tags and lemmas. Because the two mechanisms push in opposite directions, the net distortion depends on text composition and cannot be corrected by a single adjustment, which makes cross-linguistic comparison unreliable unless these decisions are made explicit.

The question of tokenization principles, what should count as a token in a language whose orthographic words align with syntactic words far less consistently than in English, has direct implications for research that measures lexical development or competence from Croatian corpora, particularly in educational contexts. Kilgarriff et al. (2014) already warned about tokenization limitations for non-English languages. The issue is not that tokenizers make errors, since tagging of these forms is consistent to within a hundredth of a percent, but that whitespace-based tokenization and the grammatical categories of English are built into the design of the pipelines, making them structurally ill-suited to languages with different morphological profiles. Corpus tools and the pipelines that feed them need clearer tokenization protocols for Slavic languages.

Keywords: lexical density, Croatian, tokenization, CLASSLA-Stanza, Sketch Engine.