Appendix

This appendix provides additional technical details supporting the Ento-Linguistic analysis presented in the main manuscript.

A. Text Processing Implementation Details

A.1 Linguistic Preprocessing Pipeline

Our text processing pipeline implements systematic normalization to ensure reliable pattern detection across diverse scientific writing styles:

$$T_{\text{processed}} = \text{lemmatize}(\text{pos}_f ilter \text{pos}_f ilter \text{pos}_f ilter_{\text{pos}_f ilter(\text{tokenize}(\text{lowercase}(}$$

where each transformation step preserves semantic content while standardizing linguistic variation.

Tokenization Strategy: We employ domain-aware tokenization that recognizes scientific terminology:

$$\tau_{\text{scientific}}(T) = \begin{cases} \text{scientific\_term}(t) & \text{if } t \in \mathcal{T}_{\text{domain}} \\ \text{word\_tokenize}(t) & \text{otherwise} \end{cases} \quad (2)$$