Language detection using character n-gram profiles

Inspiration from Cavnar and Trenkle (1994)

Atreya Shankar
Applying for: Scientific Researcher in NLP

July 6, 2021

Overview

Motivation

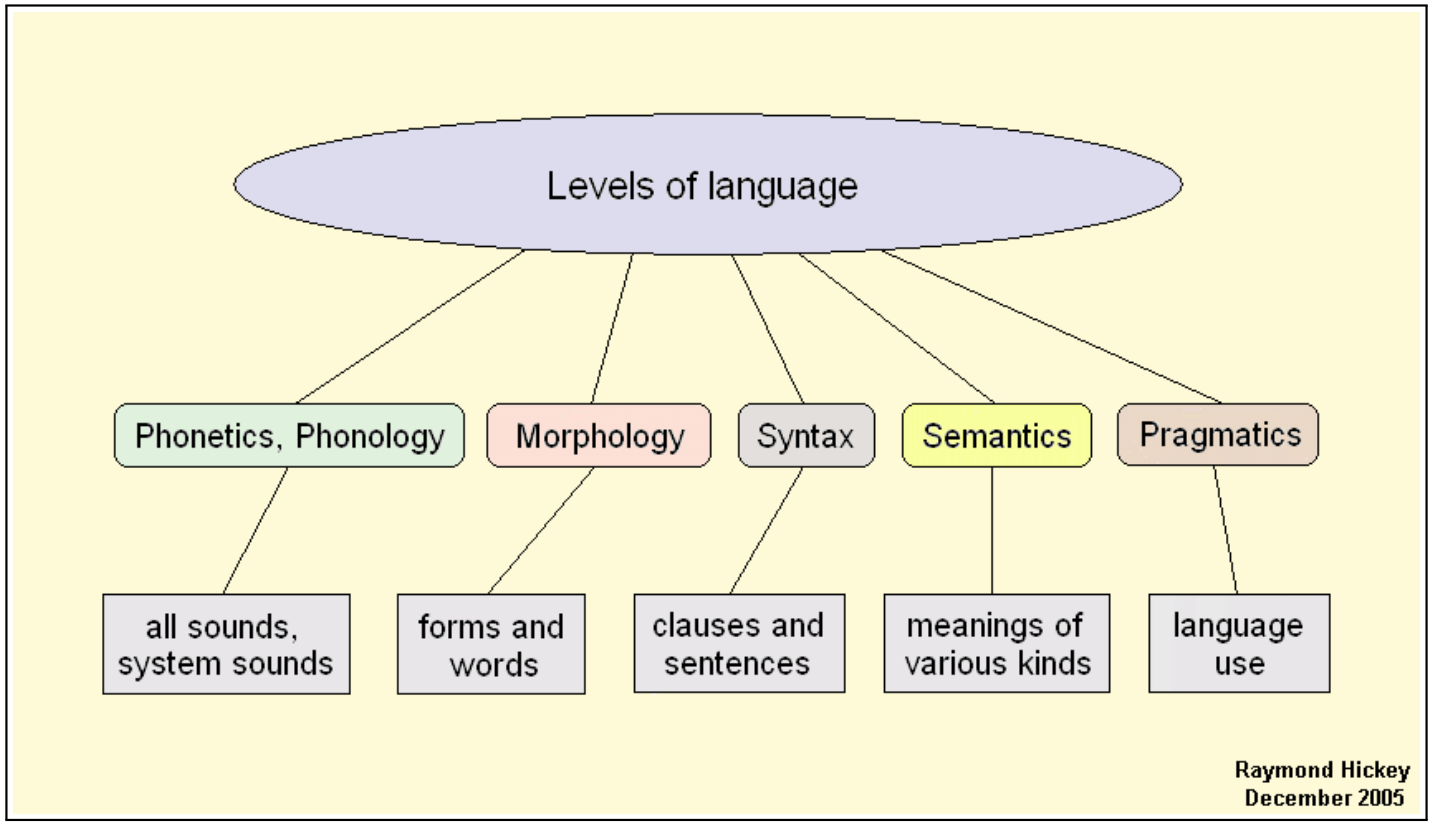

Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Motivation

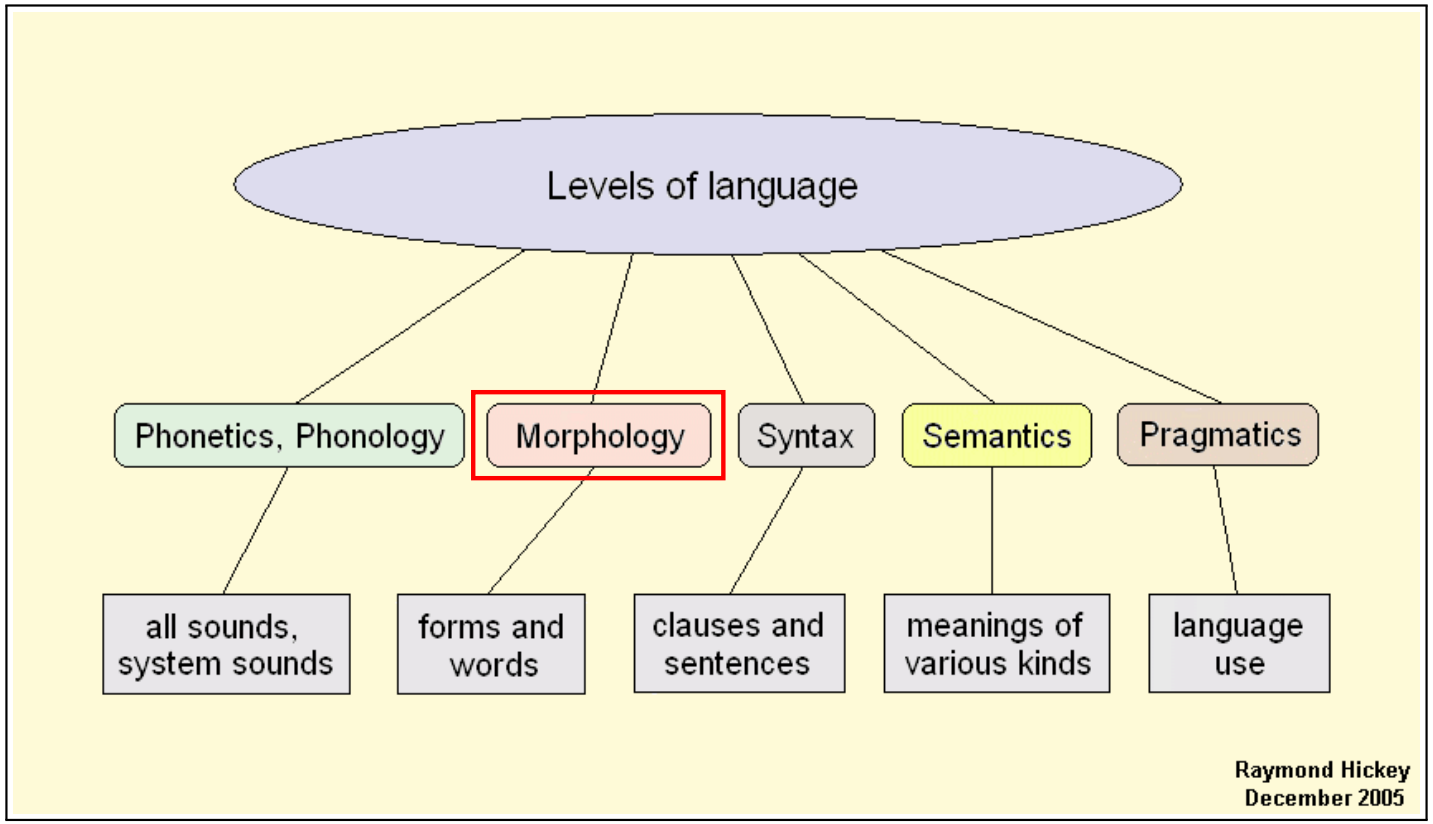

Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Motivation

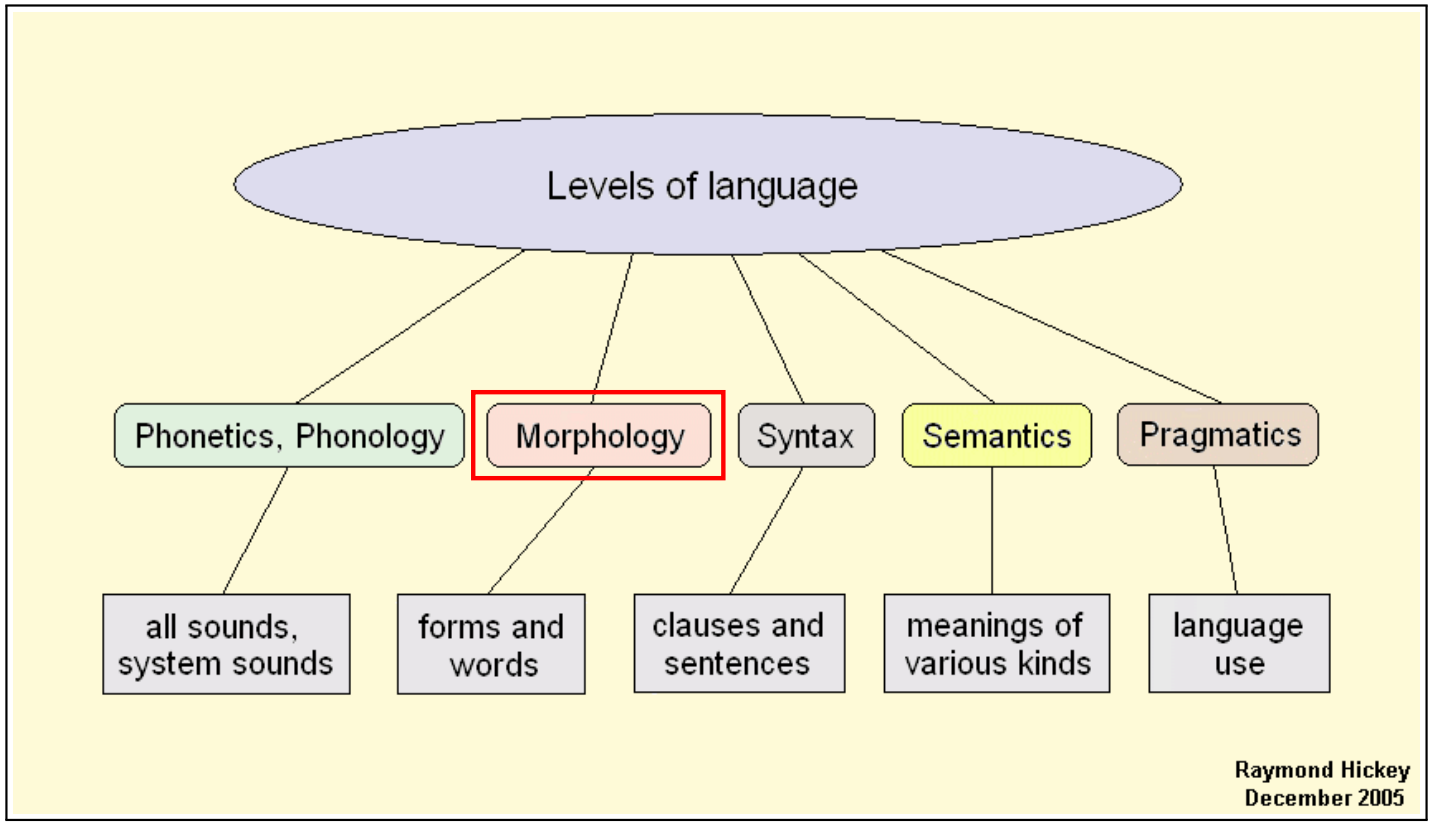

Figure 1: Levels/structures of languages; figure taken from Hickey (2005)

- Morphological profiling probably has lower data and compute requirements
- Makes sense given no external libraries are allowed

Methodology

Results

Discussion

Conclusions

Bibliography I

Cavnar, William B, John M Trenkle, et al. (1994). "N-gram-based text categorization". In: *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175. Citeseer.

Hickey, Raymond (2005). "Levels of language". In: *Universität Duisburg-Essen*.

Thoma, Martin (2018). "The WiLI benchmark dataset for written language identification". In: *arXiv preprint arXiv:1801.07779*.