Creating tools for Language is the goal

Article as the example project arc

Parsing the Dictionary

Readlines Comprehension to files

DOCX features

Dict/JSON structure

Weak for relational features and difficult to inspect/audit/modularize the parsing code

Data-structures with OOP

Dataclasses, control internal states, but still struggle with relational without a lot of extra structures

Good for Modeling entries, and modeling Paragraph objects (largest DOCX structure), but the step from one to another is difficult. Enforcing types and structure was the safest way to handle arbitrary data with limited ability to manually audit the

Enter Pydantic

Abstracting and Generalizing

Pandas as API

Networkx as API

Alternative approaches

Neo4j XML parser

MS Word Convert to HTML, and then Beautiful Soup