Zawilinski is a Java library for extracting grammar data from Wiktionary XML dumps
• Our primary use is to extract and study inflection (i.e., word ending) data.

Goals:
• Utilize existing XML parsing libraries
• Minimize code needed outside parsing libraries
• Minimize new code needed to extract new data

Main challenge:
• SAX (stream) parser efficient, but leaves a lot of work for programmer; solution doesn’t generalize
• DOM/JAXB (document) parser leaves much less work for programmer, but too memory-intensive

Our solution: Use SAX filter to remove enough data to make DOM/JAXB parsing feasible
• Much of XML dump is irrelevant for any particular study (e.g., words not part of language under study).
• Some of this irrelevant data is easy to remove as the XML streams by (see below).
• Other data cannot easily be deemed irrelevant until entire page has streamed by (see below).
• Fortunately, removing only “obviously” irrelevant data is enough to allow JAXB to work efficiently.

Wiktionary XML dump

Sample XML data

SAX “pre”-filter
Retains all meta-data
• Small compared to size of article text
• Don’t know whether to keep until text analyzed
Discards text unrelated to language under study
• Unnecessary text easy to identify and remove
• Most text is unnecessary
Retains text related to language under study

Inflection data
• Placed into Java objects
• Objects mapped to base word and revision

Unmarshal and post-filter
• Creates Java object for each XML element
• Discards objects without inflection data

Key Benefits:
• Analyzing changes in inflection data over time requires only 300 lines of Java code in addition to Zawilinski and template expansion
• Extracting new data is simply a matter of writing new pre- and/or post- filters, then extracting desired data from article text