Zawilinski: A Library for Studying Grammar in Wiktionary

Zachary Kurmas Grand Valley State University

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:

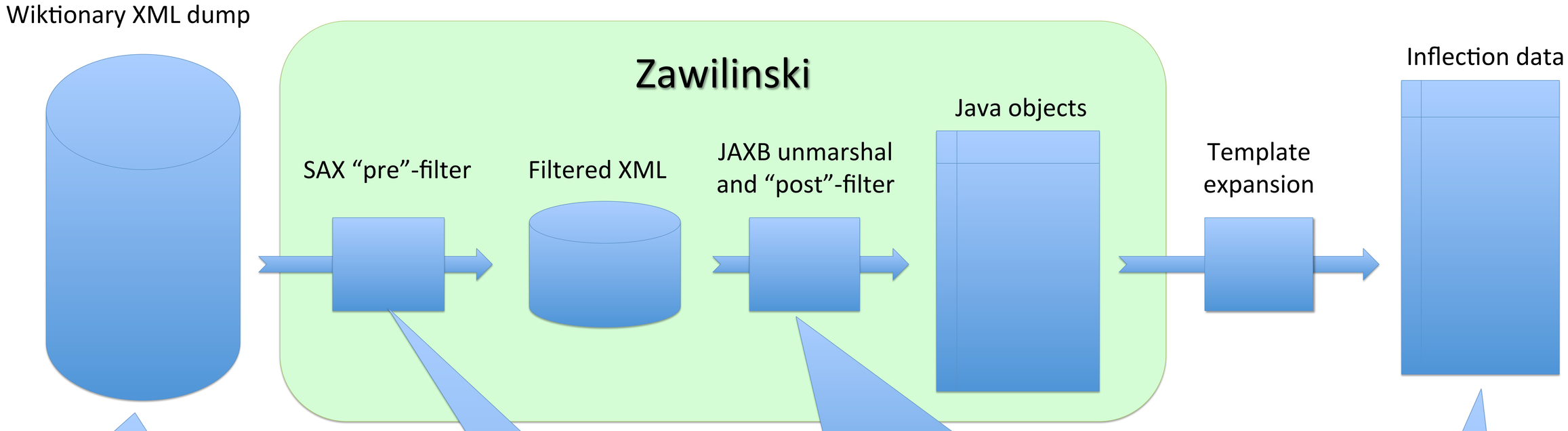
- Utilize existing XML parsing libraries
- Minimize code needed outside parsing libraries
- Minimize new code needed to extract new data

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.



Sample XML data

```
<mediawiki>
 <page>
  <title>pies</title><id>57</id>
  <revision>
   <id>7772129</id>
   <timestamp>2009-11-18T20:44:27Z</timestamp>
   <contributor>
    <username>Conrad</username><id>345</id>
   </contributor>
   <text xml:space="preserve">
==English==
===Noun===
'''pies'''# {{plural of|[[pie]]}}
...
==Polish==
===Etymology===
From {{proto|Slavic|pьsъ|lang=pl}}.
===Pronunciation===* {{IPA|/ˈpʲɛs/|lang=pl}}*
 {{audio|Pl-pies.ogg|Audio}}
===Noun===
{{infl|pl|noun|gender=m|animate}}# [[dog]]
====Declension====
{{pl-decl-noun|pies|psy|psa|psów|psu|psom|psa|
  psy|psem|psami|psie|psach|psie|psy}}
...
==Spanish==
...
```

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

Unmarshal and post-filter

- Creates Java object for each XML element
- Discards objects without inflection data

Inflection Data

- Placed into Java objects
- Objects mapped to base word and revision

Pies
2009-11-18T20:44:27Z

Nom. Sing.	pies
Gen. Sing.	psa
Dat. sing.	psu
Acc. Sing.	psa
Inst. Sing.	psem
Loc. Sing.	psie
Voc. Sing	psie
Nom. Pl.	psy
Gen. Pl.	psów
Dat. Pl.	psom
Acc. Pl.	psy
Inst. Pl.	psami
Voc. Pl.	psy

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