Voces
R-based Dashboard for Lexical Semantics

Krzysztof Nowak

Lexicon Mediae et Infimae Latinitatis Polonorum
Institute of Polish Language
Polish Academy of Sciences

8/10/2017
Overview

Why?

Motivation

- domain: lexical studies (use case: Medieval Latin lexicology)
- goals
  - quick inspection of word’s properties
  - integration of R code for flexible presentation
  - searching for best model
- target: between fully fledged corpus query systems and *ad-hoc* scripts

Dashboards (Few 2013)

- information at a glance
- visual display
- specific objectives
- fits on a single screen
Overview

General Work Flow

LEMMA → search form → OVERVIEW → meaning → distribution

CORPUS → overview → architecture

Krzysztof Nowak (LMILP)
Overview

General Lemma View

- Sidebar
- Wordform distribution table & barplot
- Cumulative frequency
- Similar terms
- Search params
- Raw co-occurrences
In-depth Lemma View

- **Word distribution by authors and texts**
- **Basic prose description of the word distribution**
- **Wordform frequency by corpus position**

**Frequency of the lemma in the entire corpus is 472 (i.e. 472 per million words).**

The word is the 288th most frequent word in the entire corpus. In terms of frequency it is slightly less popular than paulus, constiitus, spiritaliis, sacri, quippe, and more popular than dignus, urgo, puto, item, eo.$\text{et, qui, i, sum, in.}$

**Understanding**

**Interpreting**

**Reusing**
How does it work: architecture
How does it work: R code

wrap
- process AJAX calls
- dispatch call to a relevant function
- return results back

import
- read output of CQS (implemented: NoSKetchEngine, to do: CWB)
- prepare R objects for further processing

prepare data and plots
- possibly simple
- operate mostly on text.tables

Typical scenario
1. Ajax call to OCPU
2. OCPU calls wrapper fn
3. wrapper fn dispatches call to:
   - data fn or
   - plotting fn
4. data / plotting fn operates on read-in data.table or calls import fn
5. data / plot are sent back to OCPU
6. OCPU serializes output as JSON string
analytic interface for the Corpus of Polish Medieval Latin
- handling user input
- storing search routines
- going beyond
  - two- and more lemma comparison
  - diachronic analysis
- optimization
- towards model of development

Questions?

krzysztof.nowak@ijp.pan.pl
Department of Medieval Latin
Institute of Polish Language, Kraków
(Polish Academy of Sciences)