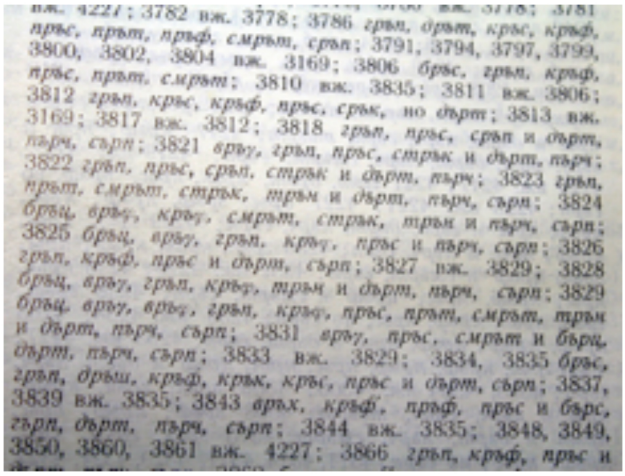
An XML-based approach to dialectological data: The development of syllabic liquids in Bulgarian

Quinn Dombrowski (quinnd@uchicago.edu), Andrew Dombrowski (adombrow@uchicago.edu)
17th Balkan and South Slavic Conference, 16 April 2010

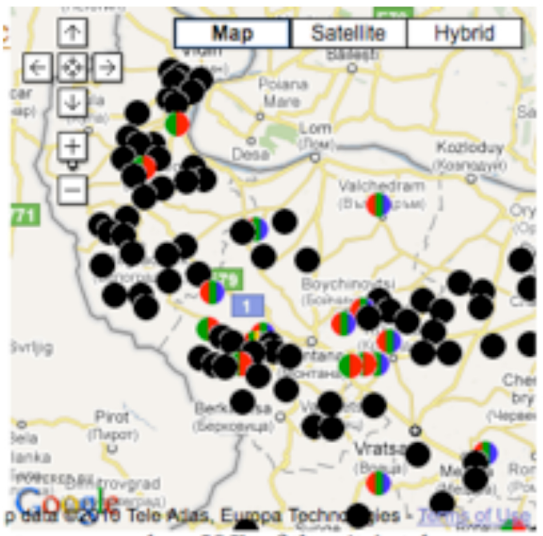
From printed text to XML



```
<site loc="NW">
   <site_number>655</site_number>
   <site_location>
      <longitude>23.349365</longitude>
      <latitude>43.387262</latitude>
   </site_location>
   <site_name>Сту`бел</site_name>
   <site_region>Михайловградско</site_region>
   <map>
      <token trt="ръ" lnum="5">гръп</token>
      <token trt="ръ" lnum="9">крък</token>
      <token trt="ръ" lnum="13">кръф</token>
      <token trt="ръ" lnum="16">пръс</token>
   </map>
</site>
```

site = each site in the atlas
@loc = region (ie, atlas volume)
  site_number = standard site number used in the atlas
  site_location = container for longitude and latitude
    longitude = longitude of site
    latitude = latitude of site
  site_name = name of site
  site_region = region of site
  map = container for tokens
    token = the word as printed in the atlas
@trt = the TrT value for the token
@lnum = a standard number created for the atlas to represent the lexeme

Flexible maps



Tool features

Site-based
- List of all sites, with an overview of which reflexes are represented in each

For each site:
- Which lexemes appear with multiple reflexes
- List of all reflexes and tokens found there

Reflexes
- List of reflexes, sorted alphabetically and by amount of sites with that reflex

For each reflex:
- How many tokens (and unique tokens) have the reflex
- How many (and what percent of) sites have the reflex
- List of sites with the reflex
- What other reflexes co-occur with the reflex

Lexeme-based
- List of lexemes, sorted alphabetically, by percent of sites with the lexeme, and by number of reflexes that occur with the lexeme

For each lexeme:
- How many (and what percent of) sites have the lexeme; list of sites can be toggled
- How many tokens (and what percent of all tokens) are from the lexeme; list of tokens can be toggled
- How many, and which, reflexes occur with the lexeme
- For which sites does this lexeme have a unique reflex?

Token-based
- List of tokens, sorted alphabetically and by number of occurrences

For each token:
- Number (and percent of) sites with the token
- Percent of lexeme instances represented by the token
- How many sites with the token have an additional form of the same lexeme
- For which sites does this token have a unique reflex?

33. Рӯпци

Located in Видинско (NW).

All	Mono	Poly
ꙋ-69.2%	ꙋ- 69.2%	ꙋ- 69.2%
ру̯-7.7%	ру̯- 7.7%	ръ- 23.1%
ръ-23.1%		

Lexeme duplicates
- кРв - ꙋ - крф
- кРв - ру̯ - кру̯ф
- кРв - ръ - кръвно

Tokens
ꙋ - 69.2%
- кр̀сти
- пр̀шки
- крф
- прс

твРд

68 sites have твРд (5.4% of sites)

68 instances of твРд (.7% of all words)

4 reflexes occur with твРд (LVC .059)

1. ꙋ - 31 tokens (1 unique)
2. ър - 24 tokens (3 unique)
3. ръ - 12 tokens (1 unique)
4. ꙋ - 1 tokens (1 unique)

Sites where твРд carries a unique тРт value

2098. Дӯшинци
- твърд - unique source of ър
- Other:
  - ръ (2)
  - ꙋ (1)

For more information about the technology

XML
A gentle introduction to XML: http://www.tei-c.org/release/doc/tei-p4-doc/html/SG.html

XSLT
W3C recommendation for XSLT 2.0: http://www.w3.org/TR/xslt20/

Cocoon
???????????