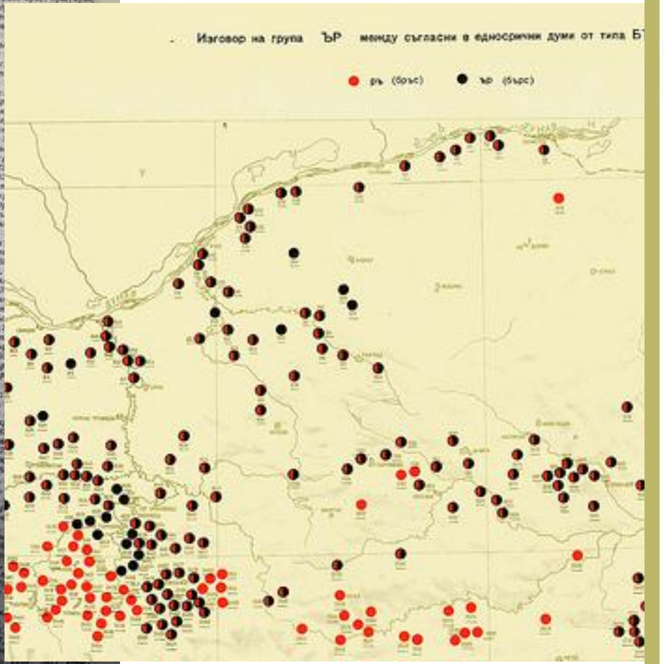
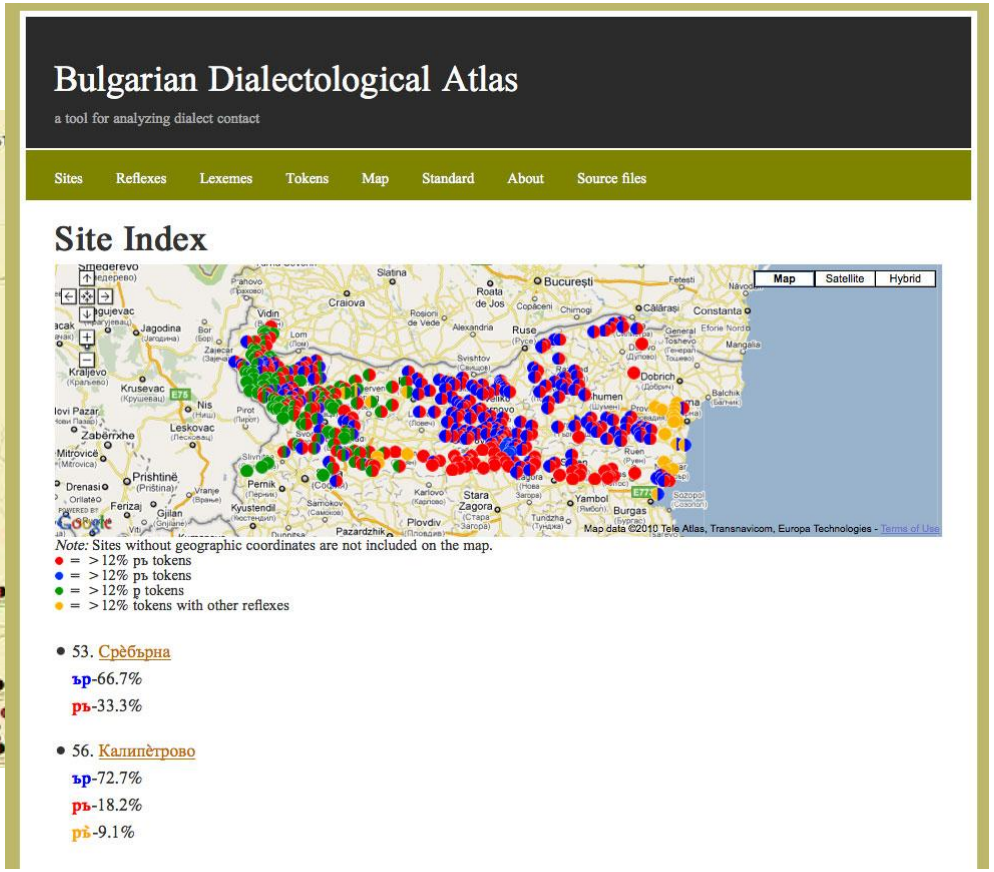
An XML-based approach to dialectological data: The development of syllabic liquids in Bulgarian


Изговор на група ЪР между съгласни в едносрични думи от типа Б
● ръ (бръс)
● ър (бърс)



Site Index

Note: Sites without geographic coordinates are not included on the map.

🔴 = >12% ръ tokens
🔵 = >12% ръ tokens
🟢 = >12% р̥ tokens
🟠 = >12% tokens with other reflexes

- 53. Срѐбърна

  ър-66.7%

  ръ-33.3%

- 56. Калипѐтрово

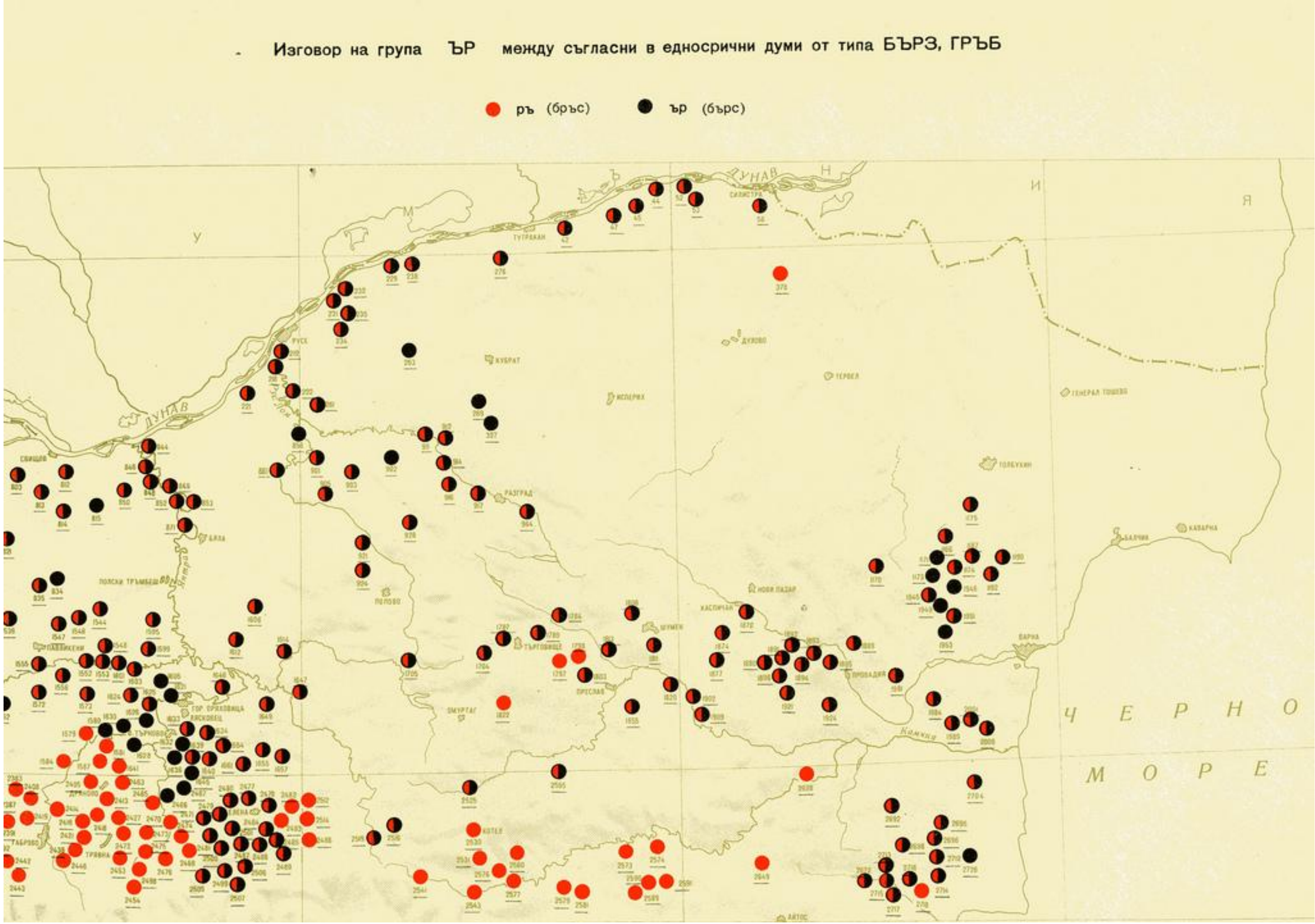
  ър-72.7%

  ръ-18.2%

  ръ̀-9.1%

Quinn & Andrew Dombrowski

To what extent do the prosodic analyses of TrT groups in standard Bulgarian characterize the dialects of Bulgaria?


Изговор на група ЪР между съгласни в едносрични думи от типа БЪРЗ, ГРЪБ
● ръ (бръс)
● ър (бърс)
ДУНАВ
ЧЕРНО МОРЕ

Sub-questions

- How many* dialects may have the pattern of behavior of the literary language?

** As determined by available data*

Sub-questions

- How many* dialects may have the pattern of behavior of the literary language?
- For those dialects that do not parallel the standard language, which of the following possibilities hold:

** As determined by available data*

Sub-questions

- How many* dialects may have the pattern of behavior of the literary language?
- For those dialects that do not parallel the standard language, which of the following possibilities hold:
  1. The distribution of TrT reflexes is purely lexical

** As determined by available data*

Sub-questions

- How many* dialects may have the pattern of behavior of the literary language?
- For those dialects that do not parallel the standard language, which of the following possibilities hold:
  1. The distribution of TrT reflexes is purely lexical
  2. The distribution of TrT reflexes is characterized by well-definable phonological conditions (not equal to those of the standard language)

* *As determined by available data*

Sub-questions

- How many* dialects may have the pattern of behavior of the literary language?
- For those dialects that do not parallel the standard language, which of the following possibilities hold:
  1. The distribution of TrT reflexes is purely lexical
  2. The distribution of TrT reflexes is characterized by well-definable phonological conditions (not equal to those of the standard language)
  3. The distribution of TrT reflexes mostly follows a regular distribution with the intrusion of discordant lexemes

* *As determined by available data*

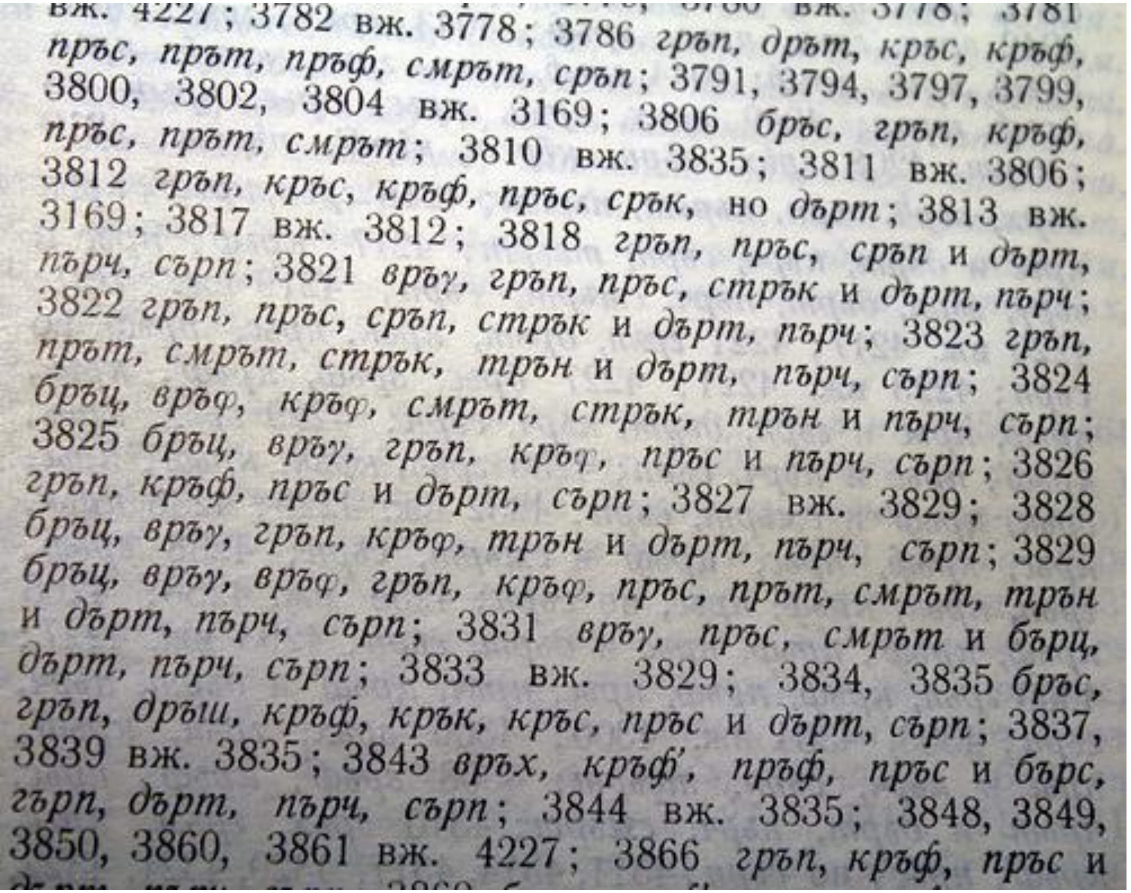
Sub-questions

- What is the role and nature of lexical diffusion in this process?
  - Just to clarify...by lexical diffusion we do not mean a non-Neogrammarian sound change.
  - Chronology:
    1. Sound change(s).
    2. Diffusion of tokens bearing various reflexes.

Why XML?

- Bulgarian Dialect Atlas (BDA) contains a *lot* of information pertaining to this...possibly too much (at first glance)!
  - Raw data lists are extremely difficult to process.
  - Maps are helpful, but impressionistic.
- XML (Extensible Markup Language) allows bottom-up rebuilding of the data set.
  - Instead of just word lists, data can be sorted and counted according to various criteria.
  - Maps can be regenerated to reflect various ways of sorting the data.

Printed edition vs. XML



3780 вж. 3778; 3781
вж. 4227; 3782 вж. 3778; 3786 *гръп, дрът, кръс, кръф,*
пръс, прът, пръф, смрът, сръп; 3791, 3794, 3797, 3799,
3800, 3802, 3804 вж. 3169; 3806 *бръс, гръп, кръф,*
пръс, прът, смрът; 3810 вж. 3835; 3811 вж. 3806;
3812 *гръп, кръс, кръф, пръс, срък,* но *дърт*; 3813 вж.
3169; 3817 вж. 3812; 3818 *гръп, пръс, сръп* и *дърт,*
пърч, сърп; 3821 *връγ, гръп, пръс, стрък* и *дърт, пърч*;
3822 *гръп, пръс, сръп, стрък* и *дърт, пърч*; 3823 *гръп,*
прът, смрът, стрък, трън и *дърт, пърч, сърп*; 3824
бръц, връφ, кръφ, смрът, стрък, трън и *пърч, сърп*;
3825 *бръц, връγ, гръп, кръφ, пръс* и *пърч, сърп*; 3826
гръп, кръф, пръс и *дърт, сърп*; 3827 вж. 3829; 3828
бръц, връγ, гръп, кръφ, трън и *дърт, пърч, сърп*; 3829
бръц, връγ, връφ, гръп, кръφ, пръс, прът, смрът, трън
и *дърт, пърч, сърп*; 3831 *връγ, пръс, смрът* и *бърц,*
дърт, пърч, сърп; 3833 вж. 3829; 3834, 3835 *бръс,*
гръп, дръш, кръф, крък, кръс, пръс и *дърт, сърп*; 3837,
3839 вж. 3835; 3843 *връх, кръф', пръф, пръс* и *бърс,*
гърп, дърт, пърч, сърп; 3844 вж. 3835; 3848, 3849,
3850, 3860, 3861 вж. 4227; 3866 *гръп, кръф, пръс* и

```
<site loc="NW">
    <site_number>655</site_number>
    <site_location>
        <longitude>23.349365</longitude>
        <latitude>43.387262</latitude>
    </site_location>
    <site_name>Сту̀бел</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>
        <token trt="ръ" lnum="35">чръф</token>
        <token trt="р̥" lnum="5">гр̥п</token>
        <token trt="р̥" lnum="16">пр̥с</token>
        <token trt="ър" lnum="20">сърп</token>
   </map>
</site>
```

Atlas data in XML

```
 <site loc="NW">
    <site_number>655</site_number>
    <site_location>
       <longitude>23.349365</longitude>
       <latitude>43.387262</latitude>
    </site_location>
    <site_name>Сту̀бел</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>
       <token trt="ръ" lnum="35">чръф</token>
       <token trt="р̥" lnum="5">гр̥п</token>
       <token trt="р̥" lnum="16">пр̥с</token>
       <token trt="ър" lnum="20">сърп</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

Lexeme index in XML

```
<lexeme>
   <word>гРп</word>
   <number>5</number>
   <token trt="ар" lnum="5">гарп</token>
   <token trt="ър" lnum="5">гърп</token>
   <token trt="ръ" lnum="5">грън</token>
   <token trt="е̂р" lnum="5">ге̂рп</token>
   <token trt="а̊р" lnum="5">га̊рп</token>
</lexeme>

<lexeme>
   <word>гРс</word>
   <number>6</number>
   <token trt="ръ" lnum="6">гръс</token>
   <token trt="о̂р" lnum="6">го̂рс</token>
   <token trt="ър" lnum="6">гърс'</token>
</lexeme>
```

lexeme = container for data relevant to each underlying "word"

word = (constructed) etymology, using P to stand in for the liquid

number = standard number to identify lexemes; identical to ***@lnum*** for each token

token = the word as printed in the atlas

@trt = the TrT value for the token

Behind the scenes

XML

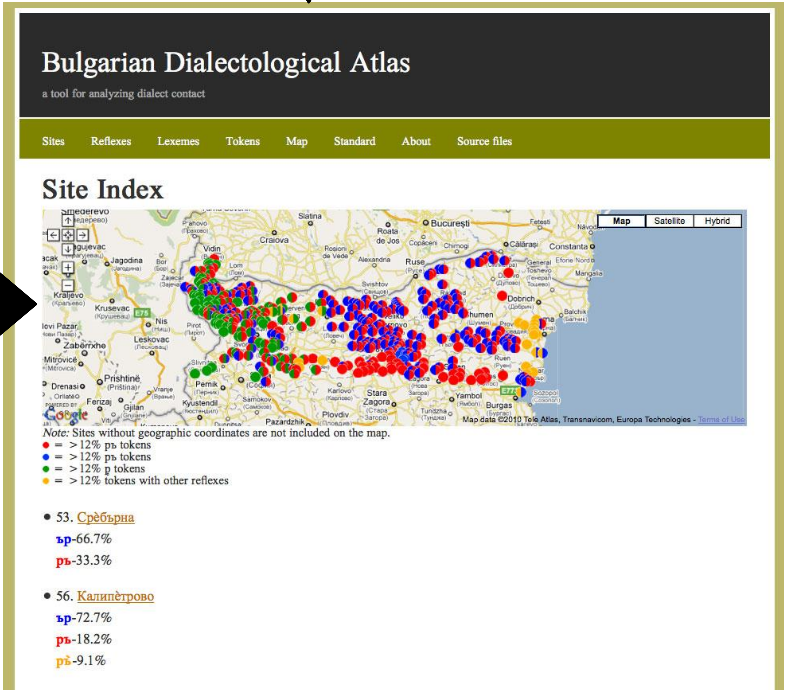
```
<atlas>
   <site>
      <site_number>9</site_number>
      <site_location>
         <longitude>22.74344</longitude>
         <latitude>44.051005</latitude>
      </site_location>
      <site_name>Плаку̀дер</site_name>
      <site_region>Видинско</site_region>
      <map mnum="107-4" data="trt1">
         <token trt="р̥" lnum="5">гр̥п</token>
         <token trt="р̥" lnum="10">кр̥с</token>
         <token trt="р̥" lnum="13">кр̥ф</token>
         <token trt="р̥" lnum="16">пр̥с</token>
         <token trt="р̥" lnum="18">пр̥ч</token>
         <token trt="р̥" lnum="20">ср̥п</token>
         <token trt="р̥" lnum="34">чр̥н</token>
      </map>
 <index>
      <lexeme>
         <word>бРс</word>
         <number>1</number>
         <token trt="ръ" lnum="1">бръс</token>
         <token trt="ър" lnum="1">бърс</token>
      </lexeme>
   </index>
</atlas>
```

+

XSLT

```
<xsl:stylesheet
xmlns:xsl="http://www.w3.org/1999/XSL/Transf
orm" version="2.0">
   <xsl:import href="site_template.xsl"/>
   <xsl:key name="aword" match="site_name"
use="../map/reflex/token"/>

<xsl:template match="atlas">
   <div id="alphabetical">
      <h3>Alphabetical</h3>
               <ul>
                  <xsl:for-each
select="index/lexeme">
                     <xsl:sort select="word"
order="ascending"/>
                     <li><a
href="lexemestats/{word}"><xsl:value-of
select="word"/></a></li>
                  </xsl:for-each>
               </ul>
      </div>
</xsl:template>
</xsl:stylesheet>
```



Bulgarian Dialectological Atlas

a tool for analyzing dialect contact

Sites Reflexes Lexemes Tokens Map Standard About Source files

Site Index

Note: Sites without geographic coordinates are not included on the map.

- 🟥 = >12% ръ tokens
- 🟦 = >12% ръ tokens
- 🟩 = >12% р̥ tokens
- 🟨 = >12% tokens with other reflexes

- 53. Срѐбърна

  ър-66.7%

  ръ-33.3%

- 56. Калипѐтрово

  ър-72.7%

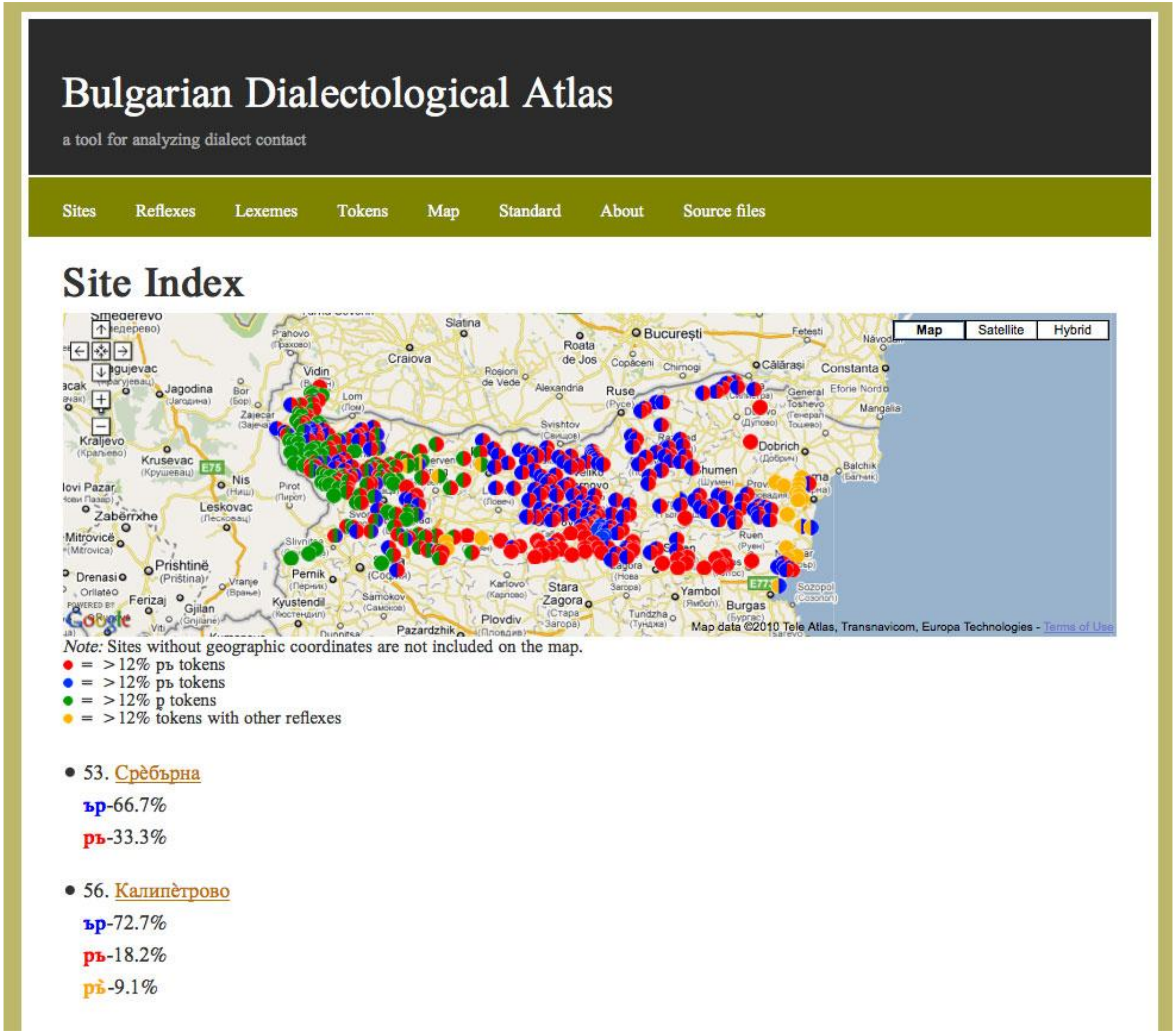
  ръ-18.2%

  рѣ-9.1%

cocoon

Apache

Site list



Bulgarian Dialectological Atlas

a tool for analyzing dialect contact

Sites Reflexes Lexemes Tokens Map Standard About Source files

Site Index

Note: Sites without geographic coordinates are not included on the map.

🔴 = >12% ръ tokens
🔵 = >12% ръ tokens
🟢 = >12% р̥ tokens
🟠 = >12% tokens with other reflexes

- 53. Срѐбърна

  ър-66.7%

  ръ-33.3%

- 56. Калипѐтрово

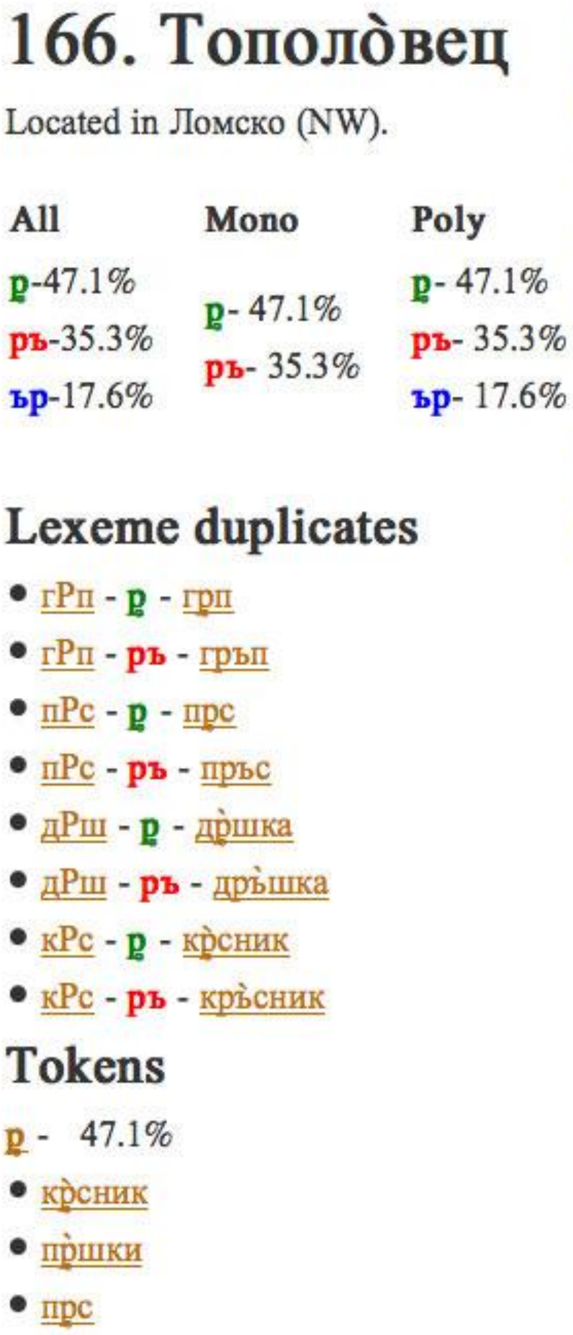
  ър-72.7%

  ръ-18.2%

  рѐ-9.1%

- List of all sites and the reflexes found there
- Map gives a visual overview of the data
- Site names are clickable to see site view

Site view



166. Тополо̀вец

Located in Ломско (NW).

All	Mono	Poly
р̥-47.1%	р̥- 47.1%	р̥- 47.1%
ръ-35.3%	ръ- 35.3%	ръ- 35.3%
ър-17.6%		ър- 17.6%

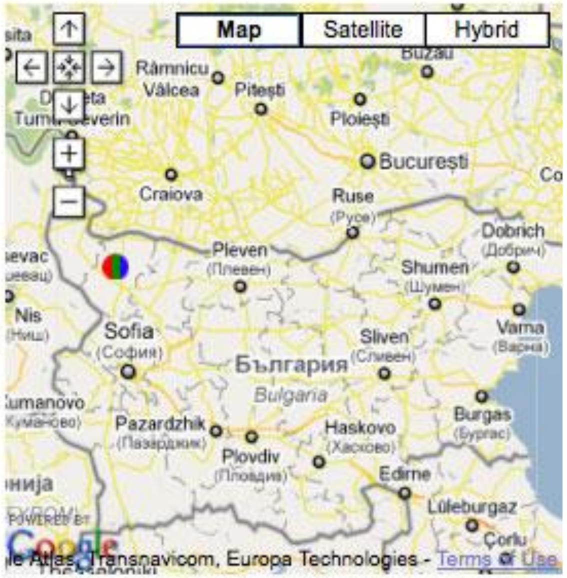
Lexeme duplicates

- гРп - р̥ - гр̥п
- гРп - ръ - гръп
- пРс - р̥ - пр̥с
- пРс - ръ - пръс
- дРш - р̥ - др̥̀шка
- дРш - ръ - дръ̀шка
- кРс - р̥ - кр̥̀сник
- кРс - ръ - кръ̀сник

Tokens

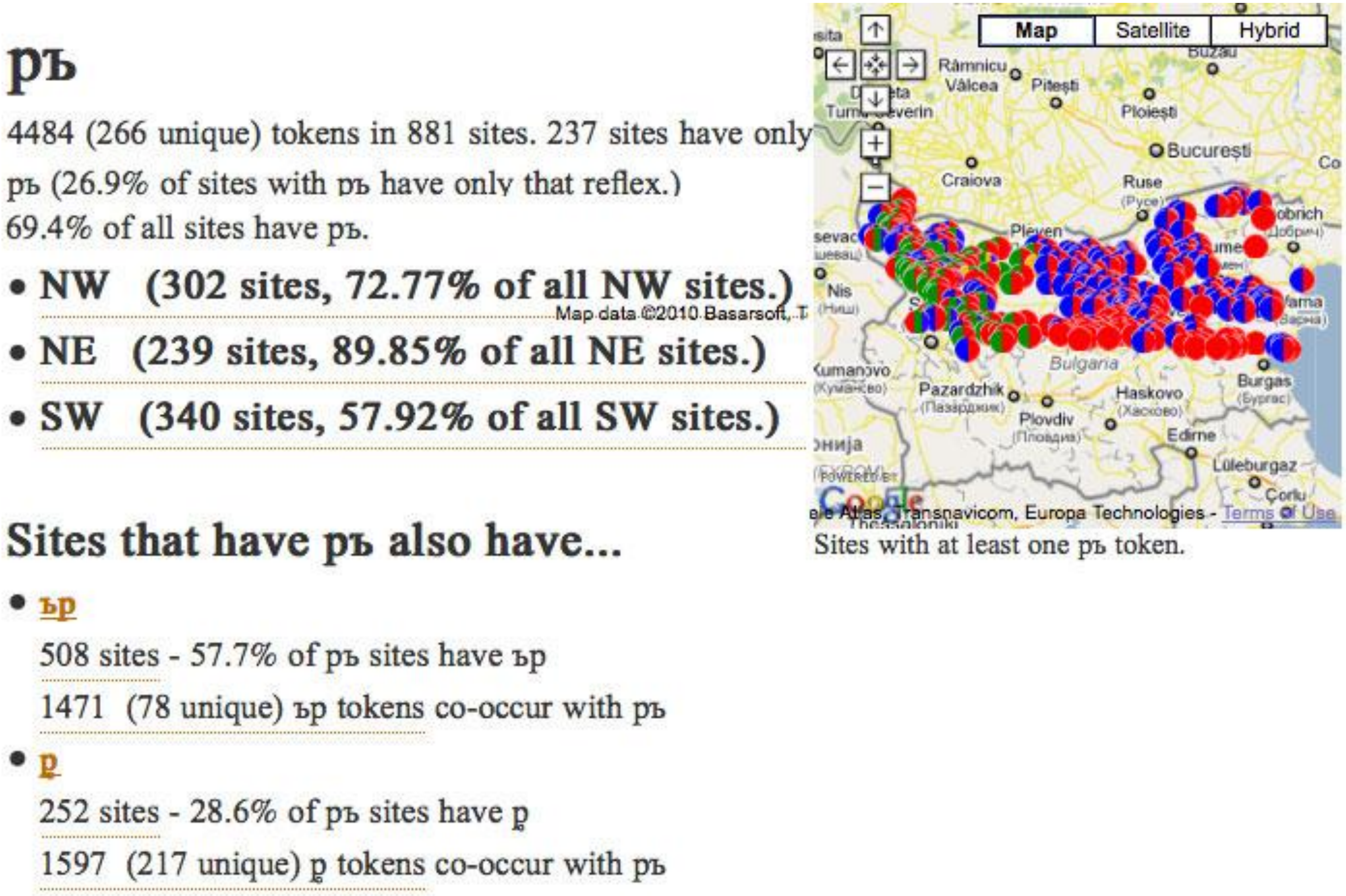
р̥ - 47.1%

- кр̥̀сник
- пр̥̀шки
- пр̥с



- Percentages are provided for each reflex found at the site
- Where a lexeme displays multiple reflexes, those lexemes and the tokens are identified; both are clickable for more detail
- A list of all tokens from the site is available; all tokens and reflexes are clickable for more detail
- A map shows the location of the site

Reflex view



pъ

4484 (266 unique) tokens in 881 sites. 237 sites have only pъ (26.9% of sites with pъ have only that reflex.)
69.4% of all sites have pъ.

- **NW (302 sites, 72.77% of all NW sites.)**
- **NE (239 sites, 89.85% of all NE sites.)**
- **SW (340 sites, 57.92% of all SW sites.)**

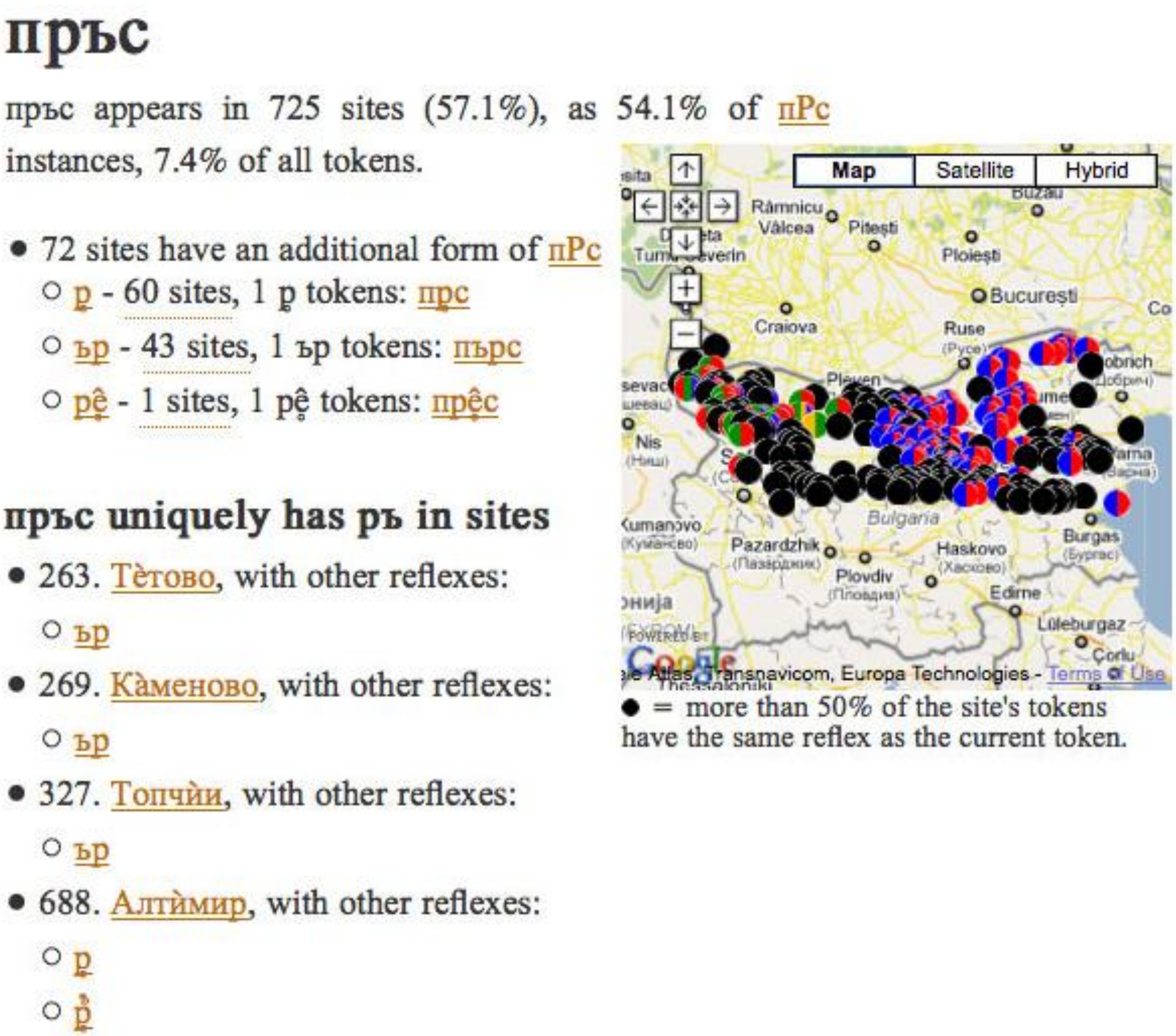
Sites with at least one pъ token.

Sites that have pъ also have...

- ър
  508 sites - 57.7% of pъ sites have ър
  1471 (78 unique) ър tokens co-occur with pъ
- р̥
  252 sites - 28.6% of pъ sites have р̥
  1597 (217 unique) р̥ tokens co-occur with pъ

- A count of all the tokens with the reflex, all the sites with the reflex, what % of all sites have the reflex, and what % of sites only have the reflex
- Toggle-down lists of sites with the reflex for each region
- What reflexes co-occur with the reflex, and with what frequency

Token view



пръс

пръс appears in 725 sites (57.1%), as 54.1% of пРс instances, 7.4% of all tokens.

- 72 sites have an additional form of пРс
  - р̥ - 60 sites, 1 р̥ tokens: пр̥с
  - ър - 43 sites, 1 ър tokens: пърс
  - рê̥ - 1 sites, 1 рê̥ tokens: прê̥с

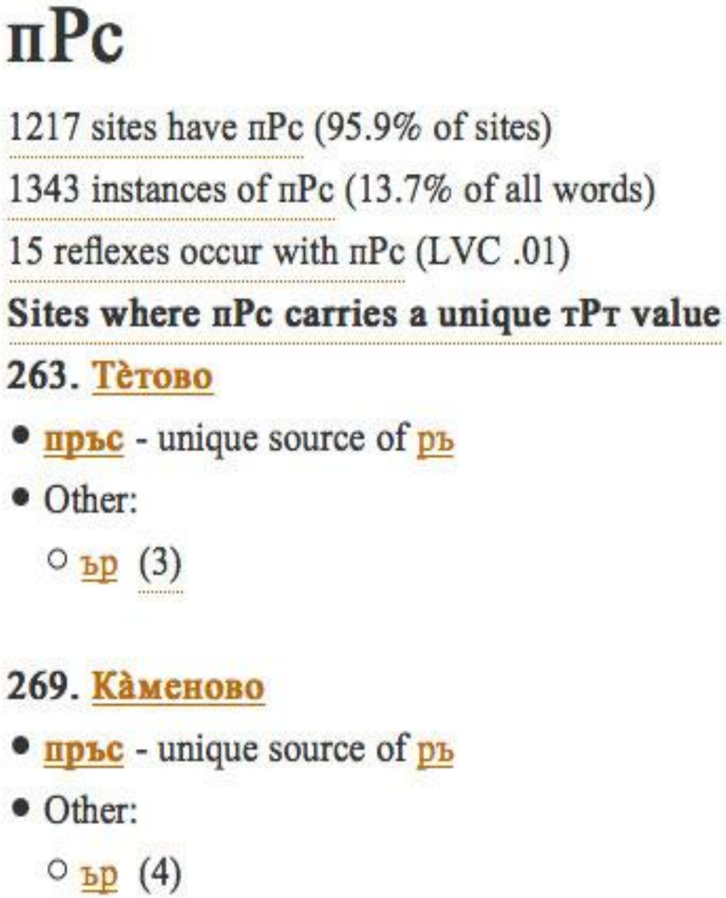
● = more than 50% of the site's tokens have the same reflex as the current token.

пръс uniquely has ръ in sites

- 263. Тѐтово, with other reflexes:
  - ър
- 269. Кàменово, with other reflexes:
  - ър
- 327. Топчѝи, with other reflexes:
  - ър
- 688. Алтѝмир, with other reflexes:
  - р̥
  - р̥ᵊ

- Lists how many sites have the token, and what % of all lexeme instances the token represents
- Lists the sites where the token is the only instance of its reflex

Lexeme view



пРс

1217 sites have пРс (95.9% of sites)

1343 instances of пРс (13.7% of all words)

15 reflexes occur with пРс (LVC .01)

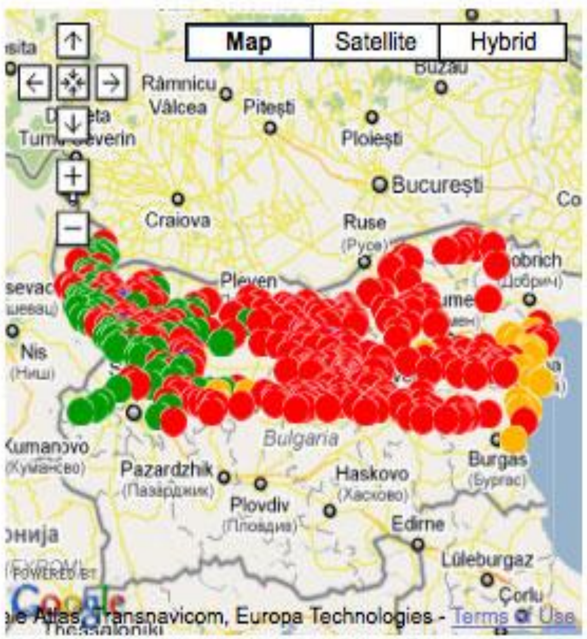
Sites where пРс carries a unique тРт value

263. Tèтово

- пръс - unique source of ръ
- Other:
  - ○ ър (3)

269. Kàменово

- пръс - unique source of ръ
- Other:
  - ○ ър (4)



Sites with пРс, shown in the reflex color of the token.

- Count of how many sites have the lexeme, how many instances there are, and how many reflexes appear with the lexeme
- A list of the relevant sites, instances, etc. can be toggled down
- List of sites where the lexeme carries a unique TrT value

How many dialects may have the pattern of behavior of the literary language?

Approximate upper bound; adding polysyllabic data and data with complex codas will reduce the number of conforming

How many dialects may have the pattern of behavior of the literary language?

Approximate upper bound; adding polysyllabic data and data with complex codas will reduce the number of conforming

12 (.9%)

Of those dialects that do not parallel the standard language, for how many is the distribution of TrT reflexes purely lexical?

Here defined as "no single reflex can be found in 75% or more of the tokens of the site".

Of those dialects that do not parallel the standard language, for how many is the distribution of TrT reflexes purely lexical?

Here defined as "no single reflex can be found in 75% or more of the tokens of the site".

471 (37%)

Of those dialects that do not parallel the standard language, for how many does the distribution of TrT reflexes is characterized by well-definable phonological conditions?

Here defined as "sites where all monosyllabic tokens carry the same reflex, excluding sites where all monosyllabic tokens carry the reflex ръ".

Of those dialects that do not parallel the standard language, for how many does the distribution of TrT reflexes is characterized by well-definable phonological conditions?

Here defined as "sites where all monosyllabic tokens carry the same reflex, excluding sites where all monosyllabic tokens carry the reflex ръ".

299 (24%)

For those dialects that do not parallel the standard language, for how many does the distribution of TrT reflexes mostly follows a regular distribution with the intrusion of discordant lexemes?

Here defined as "sites where the reflex with the most number of tokens appears in 75-99% of the tokens in that site".

Of those dialects that do not parallel the standard language, for how many does the distribution of TrT reflexes mostly follows a regular distribution with the intrusion of discordant lexemes?

Here defined as "sites where the reflex with the most number of tokens appears in 75-99% of the tokens in that site".

249 (20%)

Is lexical diffusion basically random, or do some words tend to diffuse more?

- MANY different possible metrics to get at this.
- Lexemes are attested with 1-16 discrete reflexes; what conditions this?
  - Chance: # of attested reflexes is strongly correlated with # of attested locations; r = .8568, p < .0001.
- How often are certain lexemes is the bearer of a unique tRt reflex at some geographic point?
  - # of unique tRt reflexes varies from 0 to 32.
  - # of unique tRt reflexes is strongly correlated with # of attested locations; r = .8949, p < .0001.
- Lexical diffusion seems to be basically **random**.
  - This agrees with impressionistic assessments...
  - ...but would be difficult to prove based on the atlas alone.

Conclusions

- XML markup of pre-existing data set allows a much more nuanced application that would otherwise be possible.
  - This enables answering linguistic questions that would otherwise be near-intractable.
  - Suggests ways to maximize utility of scholarly heritage.
- Problems / Future Steps:
  - Incomplete / inconsistent data across volumes.
    - e.g., "generally X, but here's some Y" for polysyllables.
  - What quantitative metrics to apply to the data?
    - Incorporation of geographic data
    - Similarity metrics to compare geographic points, the geographic distribution of reflexes, etc.
  - Research questions similar, but orthogonal to Buldialect project (Osenova et al. 2007, Heeringa et al. 2010).

References

- Barnes, Jonathan. 1997. "Bulgarian Liquid Metathesis and Syllabification in OT." in Bošković, Željko, Steven Franks, and William Snyder, eds. Annual Workshop on Formal Approachs to Slavic Linguistics: the Connecticut Meeting: 3853.
- Heeringa, Wilbert, Petya Osenova, and John Nerbonne. 2010. "Detecting Contact Effects in Pronunciation." in Hasselblatt, Cornelius, et al., eds. Language Contact: New Perspectives. Amsterdam: John Benjamins. pp. 131-153.
- Osenova, Petya, Wilbert Heeringa, and John Nerbonne. 2007. "A Quantitative Analysis of Bulgarian Dialect Pronunciation." Forthcoming in *Zeitschrift für Slavische Philologie.*
- Scatton, Ernest. 1976. “Liquids, schwa, and vowel-zero alternations in modern Bg.” in Butler, ed. Bulgaria Past and Present. Columbus: 323-327.

Sources for XML and XSLT information: on handout.