За електронната форма на средновековния български корпус от текстове и речник

Андрей Бояджиев (Софийски университет)

Хисаря, 28 февруари 2010 г.,

Преглед

- Този проект и другите инициативи в областта
- Общи изисквания
- Избор на модел
- Корпус от текстове
- Речник
- Описания: ръкописи, библиография, терминология и авторитетни файлове

Този проект на фона на други проекти

- Архиви и картотеки в България
- Titus, Манускрипт, Repertorium, Old Church Slavonic Online
- Електронни издания
- Perseus
- История и историзъм ..., SuDigital

Общи изисквания

- Unicode: 1.1.0 (Cyrillic historic letters); 4.1 (Глаголица); 5.1 (Cyrillic Extended-A, Cyrllic Extended-B);
- Единен формат на данните
- Единна конфигурация на данните

Единна конфигурация

Основни документи

текстови файлове (корпус от текстове)

речник

Помощни документи и препратки

описания на източниците (ръкописи и епиграфика); графични файлове (главно снимки от източници); библиография; терминологичен апарат; съкращения; календар със списък на светци и събития; напътствия за работа за членовете на екипа; напътствия за използване на Интернет-базираната информация от потребители; документация на проекта; технически файлове

Избор на модел

корпус от текстове ↔ речник

↑ ↑

метаданни

(описания, библиография, авторитетни файлове)

Корпуси от текстове

- Корпуси и издания
- Маркиране на езиковите характеристики
  - формални езикови правила
  - маркиране в самия корпус

Формални правила

- Нужна е изработката на норми, единни за целия корпус
- Правила от типа: “ако видиш окончание х, то въведи следната характеристика: ...”

Анотации в самия текст

- Общ модел за маркиране на езиковите данни
- <w lemma="вѣра" pos="N" gend="f" numb="sg" case="acc">вѣрѫ</w>
- <w lemma="вѣра" gram="NNfsa----------">вѣрѫ</w>

Речник

- Връзка между корпуса и речника:
  <w lemma="вѣра"
  gram="NNfsa----------">вѣрѫ</w>

```
<xsl:when test="@lemma">
<a><xsl:attribute name="href">
<xsl:text>dict.xml#</xsl:text>
</xsl:attribute>
 <xsl:apply-templates/></a></xsl:when>
```

Речникова статия

```
<entry xml:id="навадити">
      <form type="hw"><orth
xml:lang="chu">навадити</orth></form>
        <gramGrp>
         <iType value="4conj" xml:lang="chu">
          <m function="1p_sg_pres"
type="aux">ждж</m>
          <m function="2p_sg_pres"
type="aux">иши</m></iType>
          <pos>v</pos><subc>pf</subc>
        </gramGrp>
```

Речникова статия (2)

```
<sense n="1">
      <cit xml:lang="bul" type="translation">

<quote>подуча</quote><quote>подбудя</quote></ci
t>
           <cit type="translation"
xml:lang="grc">
            <quote><w xml:id="πείθω"
corresp="навадити">πείθω</w>, <w
xml:id="προβιβάζω"
corresp="навадити">προβιβάζω</w></quote>
           </cit>
```

Речникова статия (3)

```
<cit xml:lang="chu" type="example">
        <quote>архиереи же и старьци <ref
target="#πείθω">навадиша</ref> народы. да
                испросатъ варавж. їс̄а же
погоубатъ</quote>
                <bibl><rs type="mss">Codex
Sabae</rs>
                  <rs type="NT">Mt 27:20</rs>
                  <ref target="">Иванова-
Мирчева 1999:900</ref></bibl>
              </cit>
```

Данни, обслужващи корпуса и речника

- Описания на ръкописи
- Библиография
- Единна система за терминология и съкращения
- Единна система за наименования на източниците
- Съкращения на ръкописите
- Съкращения на текстове и автори
- Кодове за езици, страни и писмени системи

Данни за ръкописите

```
<listWit><head>Списък от ръкописи</head>
<witness xml:id="Supr"> <abbr
xml:lang="bul">Супр</abbr><expan>Супрасълски
сборник</expan><abbr
xml:lang="rus">Супр</abbr><expan>Супрасльская
рукопись</expan><abbr
xml:lang="lat">Supr</abbr><expan>Codex
Suprasliensis</expan>              <ptr
target="../description/Supr.xml"
type="description"/><ref type="edition"
target="../bibliography/Zaimov.xml#ZaimovJ_Kapa
ldoM1981-1982">Заимов, Капалдо 1981-
1982</ref><ref type="description"
target="../description/Supr.xml>Описание</ref><
/witness></listWit>
```

Съкращения на текстове

```
<list><item xml:id="CPG7842"><title
xml:lang="lat">Epistula ad Eusthatium</title>

<title xml:lang="bul">Послание към
Евстатий</title></item>

<item xml:id="CPG7843">

<title xml:lang="lat">Pandecta scripturae
sacrae</title>

<title xml:lang="bul">Пандекти на Светото
писмо</title>

      </item>

</list>
```

Кодове

- За обозначаване на езиците: ISO 639 (bul, grc, lat, chu)
- За имената на страните се използва стандарта ISO 3166
- За обозначаване на писмените системи се прилага стандартът ISO 15924, разработван от консорциума Unicode: кирилица (cyrl, glag)

Няколко екипа

- Корпуси от текстове
- Речникови статии
- Изготвяне на метаданни
- Координиращ екип

Благодаря за вниманието!

Хисаря, 28 февруари 2010 г.