Курсовая работа

Смык і Янка: Корпус белорусской поэзии XVIII—XX вв.

Екатерина Немкович
МЦМГН201

Структура работы

Проект разметки TEI-XML

- TEI-XML
- Уровни:
  - Библиографическая (металингвистическая)
  - Лингвистическая
  - Стиховедческая

Сбор материала корпуса

- Отбор авторов
- Скрейпинг
- Парсинг
- Систематизация

Разметка

- **Метаразметка**
  Автор, название, издание, годы создания и публикации, жанр, информация о переводе.
- Лингвистическая
  POS-tagging: часть речи, лемма, морфосинтаксическая информация.
- Стиховедческая
  Метр, размер, типы рифм, схема рифмовки, ритмические формулы стихов.

```
<titleStmt>
    <title type="main">Санет ("Замёрзла ноччу шпаркая крыніца...")</title>
    <author>
        <persName>
            <forename>Максім</forename>
            <surname>Багдановіч</surname>
        </persName>
    </author>
</titleStmt>
<sourceDesc>
    <bibl type="onlineSource">
        <name>БЕЛАРУСКАЯ ПАЛІЧКА</name>
        <idno type="url">https://knihi.com/Maksim_Bahdanovic/Saniet_Zamiorzla_noccu_sparkaj
        a_krynica.html</idno>
        <bibl type="originalSource">
            <title>Лазарук, М.А. Беларуская літаратура: вучэб. дапам. для 8-га кл. устаноў
            агульнай сярэдняй адукацыі з беларус. і рус. мовамі навучання / М.А.Лазарук,
            В.І.Русілка, І.М.Слесарава. – Мінск: Нац. ін-т адукацыі, 2011.</title>
            <date type="written">1912</date>
            <date type="print">1912</date>
        </bibl>
    </bibl>
</sourceDesc>
```

```
Замёрзла ноччу шпаркая крыніца;
<ana type="linguistic">
    <w pos="VERB" lemma="мерзнуць" msd="Past:Perf:Act:Fem:Sing">замёрзла</w>
    <w pos="ADV" lemma="ноччу">ноччу</w>
    <w pos="ADJ" lemma="шпаркая" msd="Nom:Fem:Sing">шпаркая</w>
    <w pos="NOUN" lemma="крыніца" msd="Nom:Fem:Sing:Inan">крыніца</w>
    <pc pos="PUNCT" lemma=";">;</pc>
</ana>
```

```
<formal rhyme="AbbA|AbbA|ccD|eDe" metre="iambus"/>

<lg n="1" type="quatrain" rhyme="AbbA">

    <l n="1" met="-+|-+|-+|--|-+|-/" metre="iambus" feet="5">
    Замёрзла ноччу шпаркая крыніца;

    <ana type="rhyme">
    Замёрзла ноччу шпаркая крыніца<rhyme label="A" type="fem">крыніца</rhyme>
    </ana>

    <ana type="rhythm">
        <seg type="foot">
        <seg type="syll">За</seg>
        <seg type="syll" stress="ictus">мёрз</seg>
    </seg>
        <seg type="foot">
        <seg type="syll">ла</seg>
        <seg type="syll" stress="ictus">ноч</seg>
    </ana>

</lg>
```

Разметка: открытые вопросы

- **Сосуществования разных формализмов**
  Западная традиция (TEI) и славянское стиховедение. Разные типы разметки у одного уровня.
- Многозначность терминов
  Такие слова, как метр и рифма, определяются по-разному даже у авторов в пределах одной традиции.
- Несколько уровней масштаба
  Разметка элементов от слога до строфы. Отсутствие примеров.

Сбор: отбор авторов

- Период
  Начало XVIII века — конец советского периода.
- Источник
  Анталогія беларускай паэзіі у 3 т., рэд. Р. Барадулін (1993)
- Полный список
  293 автора
- v.1
  73 автора

Сбор: выбор ресурса

Первая версия собирается из онлайн-источников.

Беларуская палічка (knihi.com):

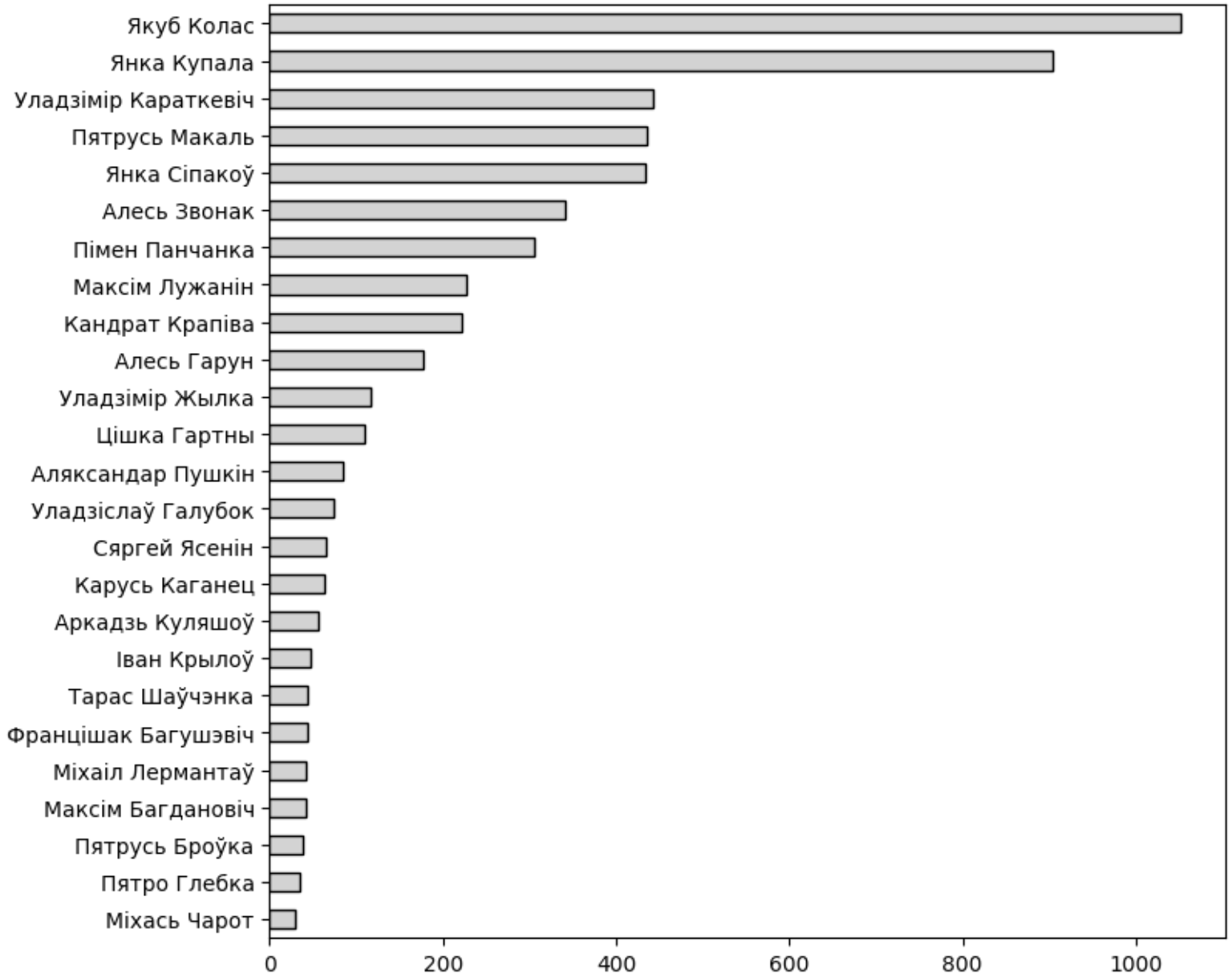
- Больше всего авторов
- Метаданные
- Простая разметка

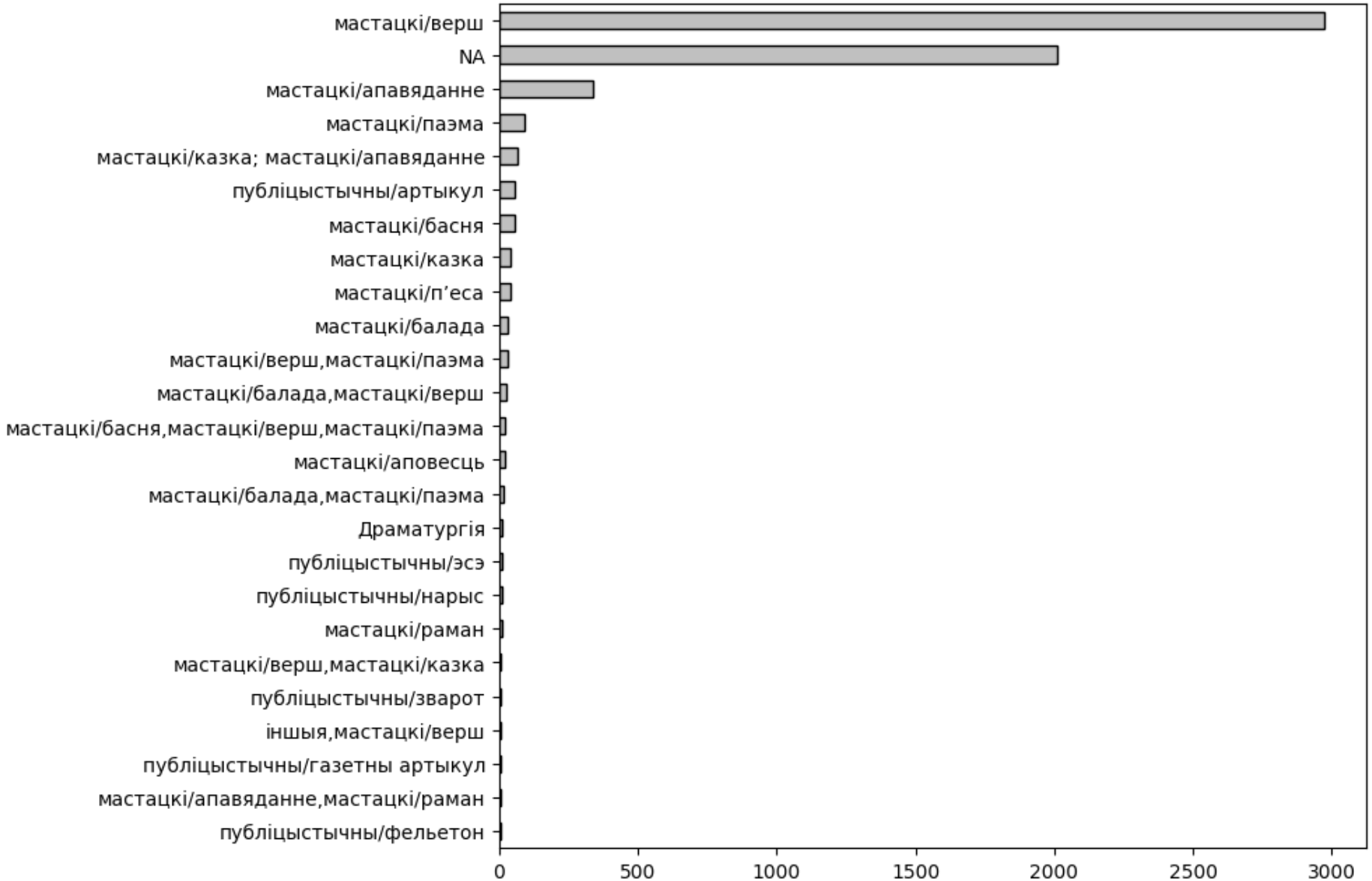
Сбор: скрейпинг и парсинг

- Сбор ссылки со страниц авторов
- Сбор произведения в текстом формате
- Парсинг их метаданных

Результаты

- 5948 текстов
- У 2012 не указан жанр
- Отмечено 2975 стихотворений, 92 поэмы, 55 басен, 31 баллада
- Есть смешанные жанры
- Есть смешанные произведения


Якуб Колас
Янка Купала
Уладзімір Караткевіч
Пятрусь Макаль
Янка Сіпакоў
Алесь Звонак
Пімен Панчанка
Максім Лужанін
Кандрат Крапіва
Алесь Гарун
Уладзімір Жылка
Цішка Гартны
Аляксандар Пушкін
Уладзіслаў Галубок
Сяргей Ясенін
Карусь Каганец
Аркадзь Куляшоў
Іван Крылоў
Тарас Шаўчэнка
Францішак Багушэвіч
Міхаіл Лермантаў
Максім Багдановіч
Пятрусь Броўка
Пятро Глебка
Міхась Чарот
0
200
400
600
800
1000


мастацкі/верш
NA
мастацкі/апавяданне
мастацкі/паэма
мастацкі/казка; мастацкі/апавяданне
публіцыстычны/артыкул
мастацкі/басня
мастацкі/казка
мастацкі/п’еса
мастацкі/балада
мастацкі/верш,мастацкі/паэма
мастацкі/балада,мастацкі/верш
мастацкі/басня,мастацкі/верш,мастацкі/паэма
мастацкі/аповесць
мастацкі/балада,мастацкі/паэма
Драматургія
публіцыстычны/эсэ
публіцыстычны/нарыс
мастацкі/раман
мастацкі/верш,мастацкі/казка
публіцыстычны/зварот
іншыя,мастацкі/верш
публіцыстычны/газетны артыкул
мастацкі/апавяданне,мастацкі/раман
публіцыстычны/фельетон
0
500
1000
1500
2000
2500
3000

Материал: открытые вопросы

- **Определение принадлежности текста к поэзии**
  Пока это корпус текстов поэтов, а не поэтический корпус.
- Полнота и репрезентативность
  Авторы представлены неравномерно.
- Качество источников
  Тексты не всегда приводятся из авторитетных источников.

Репозиторий
https://github.com/k-nem/bpcorpus