# Уеб-базирана инфраструктура за лингвистична обработка на данни на български език

## Ръководство

## Съдържание



### 1. Увод

Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове включва следните видове езикова обработка: разделяне на изречения; токънизация; автоматично определяне на частта на речта и на граматичните характеристики на думите; лематизация (автоматично определяне

на основната форма). Разработени са и някои допълнителни програми за обработка, включително за анотация и съотнасяне на паралелни текстове на ниво изречение и елементи на изречението.

Високо скалируемата инфраструктура, обединяваща отделни уеб сървиси, осигурява лесен достъп до програмите за обработка и анотация на текстове на български език. Осигурени са няколко вида достъп: онлайн достъп, REST API сървис и приложение за асинхронни задачи. Условията за достъп са изложени по-долу.

Основни характеристики на системата:

- Език: C++, PHP.
- Операционна система: Linux.

## 2. Компоненти на системата

### 2.1. Общо описание

Функцията на **Фронтенд** компонента е да прилага условията за достъп до приложно-програмния интерфейс (API) на сървисите, да се справя с грешки, идентификация на потребителите, поддържане на различни изходни формати (XML, JSON, обикновен текстов формат), комуникация с бекенд компонента. Фронтенд компонентът също така осигурява достъп до потребителски уеб интерфейс за подаване и изпълнение на асинхронни задачи: начало, спиране и наблюдение върху задачите, както и качване и сваляне на данни.

**Бекенд** компонентът от друга страна извършва реалната обработка на данните и съвместява компонените на системата - токънизатор, разделител на изречения, тагер и лематизатор, под формата на сървър приложение, което получава

и изпълнява заявките от фронтенд компонента през tcp/ip. При необходимост се прилага разпределена обработка, като работата се разпределя върху повече от една машини в мрежа, за да се увеличи ефективността при изпълнение на едновременни заявки.

**Диспечер** компонентът отговаря за управлението на процесите при асинхронните задачи. Той получава командите за начало/спиране от фронтенда и изпраща съобщение по имейл до потребителя, когато резултатът е готов.

## 2.2. Компоненти за лингвистична анотация

***Разделител на изречения и токънизатор за български език***

Разделителят на изречения и токънизаторът използват регулярни изрази. Токънизаторът разпознава съкращения, лични имена, числови изрази, дати. Разделителят на изречения и токънизаторът са интегрирани заедно с цел по-добра ефективност.

***Тагерът BgTagger за български език***

Тагерът определя най-вероятната част на речта за дадена дума в конкретен контекст и ѝ приписва еднозначна морфосинтактична информация. Тагерът е базиран на Метода на опорните вектори (Support Vector Machines) и предсказва частта на речта въз основа на множество от характеристики, които описват думата и нейния контекст.

Тези характеристики включват:

- думи, дву- и триграми от думи в контекст от няколко токъна вляво и вдясно спрямо тагираната словоформа;
- тагове за частта на речта, дву- и триграми от такива тагове в определения

контекстов прозорец;

- информация за суфикси, префикси, главни букви, сричкопренасяне и др. за думи, които не са неразпознати от речника.

Тагерът е трениран и тестван върху корпус, в който частите на речта и граматичните характеристики на думите са определени от експерти (БулПосКор, http://dcl.bas.bg/poscor/bg/). Стратегията за трениране има следните параметри: (i) две обхождания вляво и вдясно; (ii) контекстов прозорец от пет токъна, като тагираната дума е на втора позиция; (iii) дву- и триграми от думи или части на речта, лексикални параметри като префикси, суфикси, граница на изречение, главни букви и др.

Тренираният езиков модел е използван за еднозначно приписване на част на речта и морфосинтактични характеристики в корпуси от текстове на български език. Програмата има точност от 96,58%.

***Българският лематизатор***

Българският лематизатор определя основната форма на думите и ѝ преписва подробна граматична информация. Лематизацията се базира на резултата от тагирането и информацията от Граматичния речник. За тагирането се използва редуциран тагсет (75 класа в съпоставка с 1029 уникални граматични тага в речника), компилиран по начин, който осигурява минималната необходима информация за еднозначно съотнасяне със съответната лема. За разрешаване на многозначността се прилагат малък брой правила и ограничения.

## 3. Достъп

По-нататък са описани потребителският интерфейс и функционалностите, позволяващи достъп до уеб базираната инфраструктура за лингвистична обработка

на български текстове. Всички услуги са безплатни, но някои от тях изискват регистрация и идентификация при достъп. Регострация може да се направи на следния адрес:

http://dcl.bas.bg/dclservices/registration/

### 3.1. Непосредствен онлайн достъп (демо сървис)

Непосредственият достъп чрез демо приложение в интернет е подходящ за потребители, които извършват обработване на относително малки по обем данни еднократно или рядко. Приложението е достъпно на адрес: http://dcl.bas.bg/dclservices/. Броят заявки е неограничен. Резултатът е представен във вертикален анотиран формат, като всеки ред представя езиков елемент (дума, пунктуация) и неговото описание.

### 3.2. REST API

Приложно-програмният интерфейс (API) е подходящ за специалисти, които желаят да интегрират инструментите за обработка на български език в своите софтуерни разработки. Няколко типа заявки са възможни, които имат следната обща форма:

```
http://dcl.bas.bg/dclservices/service/service_name/method/return_data_format/?arg1=val1&arg2=val2&.....argn=valn
```

Аргументът `token=XXX` е задължителен и се използва за идентификация на потребителя пред системата. Доставчикът на услугата (Секцията по компютърна лингвистика, http://dcl.bas.bg/) осигурява уникален тоукън за идентификация на всеки потребител при регистрация.

По-долу са представени няколко примера за възможните API заявки.

***Токънизатор***

Заявка:

http://dcl.bas.bg/dclservices/service/bglpc/tokenizer/xml/?text=Иван%20рита%20топка&token=XXX

Резултат:

```
<DCL_SERVICE>
      <error>0</error>
      <success>true</success>
      <text>
      <item>Иван TOK_FUCA   </item>
      <item>рита TOK_LCA    </item>
      <item>топка     TOK_LCA<S> </item>
      <item/>
      </text>
</DCL_SERVICE>
```

***Тагер***

Заявка:

http://dcl.bas.bg/dclservices/service/bglpc/tagger/xml/?text=Иван%20рита%20топка&token=XXX

Резултат:

```
<DCL_SERVICE>
      <error>0</error>
      <success>true</success>
      <text>
      <item>Иван TOK_FUCA   NHs   </item>
      <item>рита TOK_LCA    Vs </item>
      <item>топка TOK_LCA<S> Ns   </item>
      <item/>
      </text>
</DCL_SERVICE>
```

***Лематизатор***

Заявка:

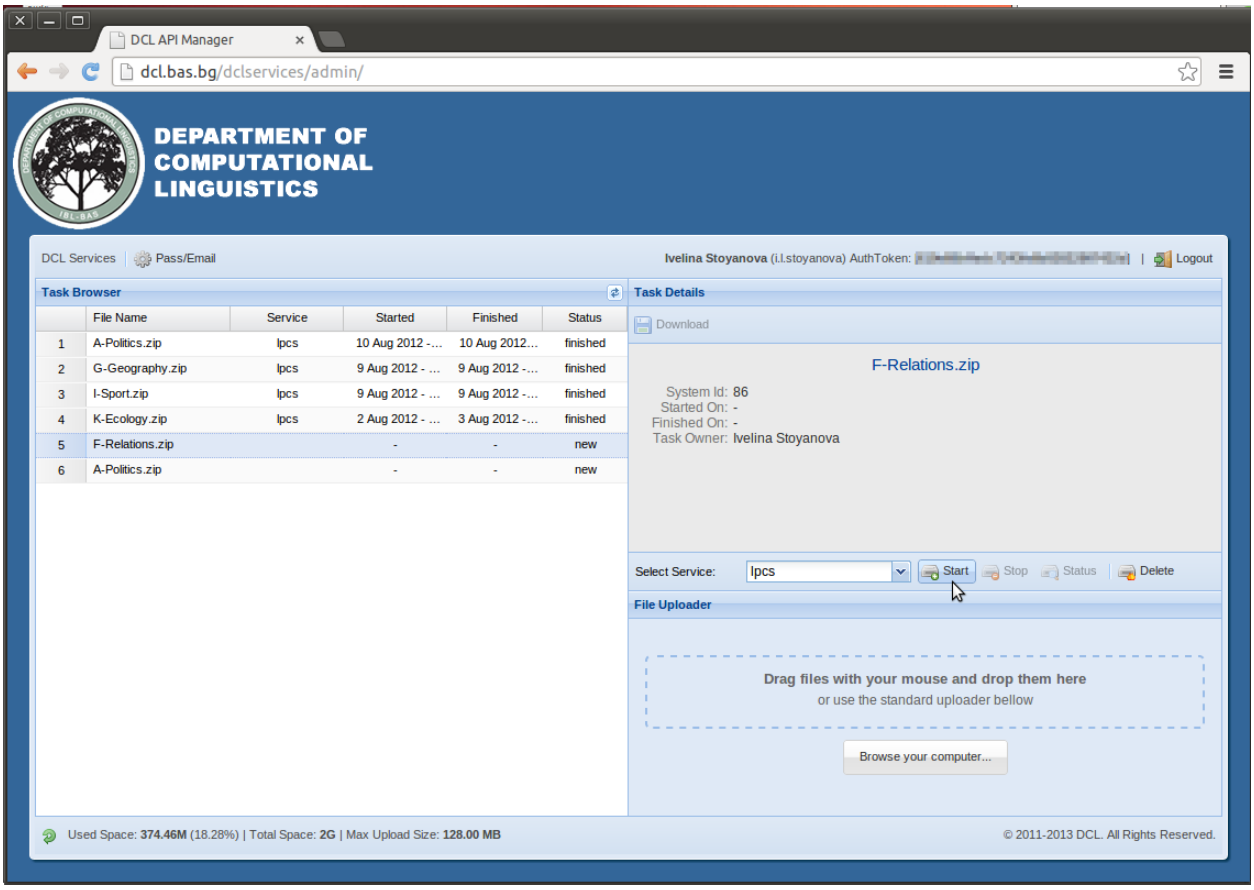
http://dcl.bas.bg/dclservices/service/bglpc/lematizer/xml/?text=Иван%20рита%20топка&token=XXX

Резултат:

```
<DCL_SERVICE>
      <error>0</error>
      <success>true</success>
      <text>
      <item>Иван TOK_FUCA   иван  NHMsom</item>
      <item>рита TOK_LCA    ритам VLITe3s</item>
      <item>топка     TOK_LCA<S> топка NCFsof</item>
      <item/>
      </text>
</DCL_SERVICE>
```

### 3.3. Асинхронни задачи

Асинхронният достъп е подходящ за задачи, изискващи по-дълго време на изпълнение, каквато е обработката на големи по обем корпуси. Изисква се регистрация за достъп до услугата. Асинхронните заявки се пускат през API мениджър за достъп до уеб интерфейса за лингвистична облаботка: http://dcl.bas.bg/dclservices/admin/.




**Фигура 1. API Мениджър на уеб интерфейса: асинхронни задачи.**

При този тип достъп потребителят предоставя за обработка архивирания корпус (в `.zip` формат) чрез интерфейса на уеб инфраструктурата, изпраща заявка за дадена анотационна задача (например `lpcs`), като при приключване на задачата системата уведомява автоматично потребителя чрез имейл, след което той може да изтегли анотирания корпус през API мениджъра.

## 4. Предимства на системата

Основните предимства на уеб инфраструктурата са следните:

- позволява висококачествена лингвистична обработка на езикови ресурси

за български език;

- осигурява комплексна и взаимно съвместима анотация на различни езикови нива;
- имплементирана с най-съвременните технологии;
- осигурява различни нива на достъп в съответствие с нуждите на различните потребители;
- високо скалируема, позволява разпределение на процесите на различни компютри.

## Източници

Koeva, S. & Genov, A. Bulgarian language processing chain. IN: *Proceedings of Integration of multilingual resources and tools in Web applications*. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg, 2011.