Bulgarian Language Technology for Digital Humanities: a Focus on the Culture of Giving for Education

Kiril Simov and Petya Osenova

CLaDA-BG

IICT-BAS

CLARIN annual meeting 2018, Pisa, Italy

Plan of the Talk

- Background
- The specialized corpus on giving
- Challenges to NLP
- Some analysis
- Conclusions

Introduction

A national project: Culture of giving in the sphere of education: social, institutional and personality dimensions at the *Institute for the Study of Societies and Knowledge* at BAS.

Two approaches:

- Application of software developed especially for the content analysis of historical documents
- Application of the theory of planned behavior to the study of philanthropy

Background

- Focus on the culture of giving for education
- Partners – from Institute of Sociology, BAS
- The collected corpus comprises texts with a time span of 80-100 years
- The task is: to extract relevant information with the help of statistics and content analysis for displaying the tendencies from the perspective of the language/phrasing/terminology, the social and economical context.

Background (2)

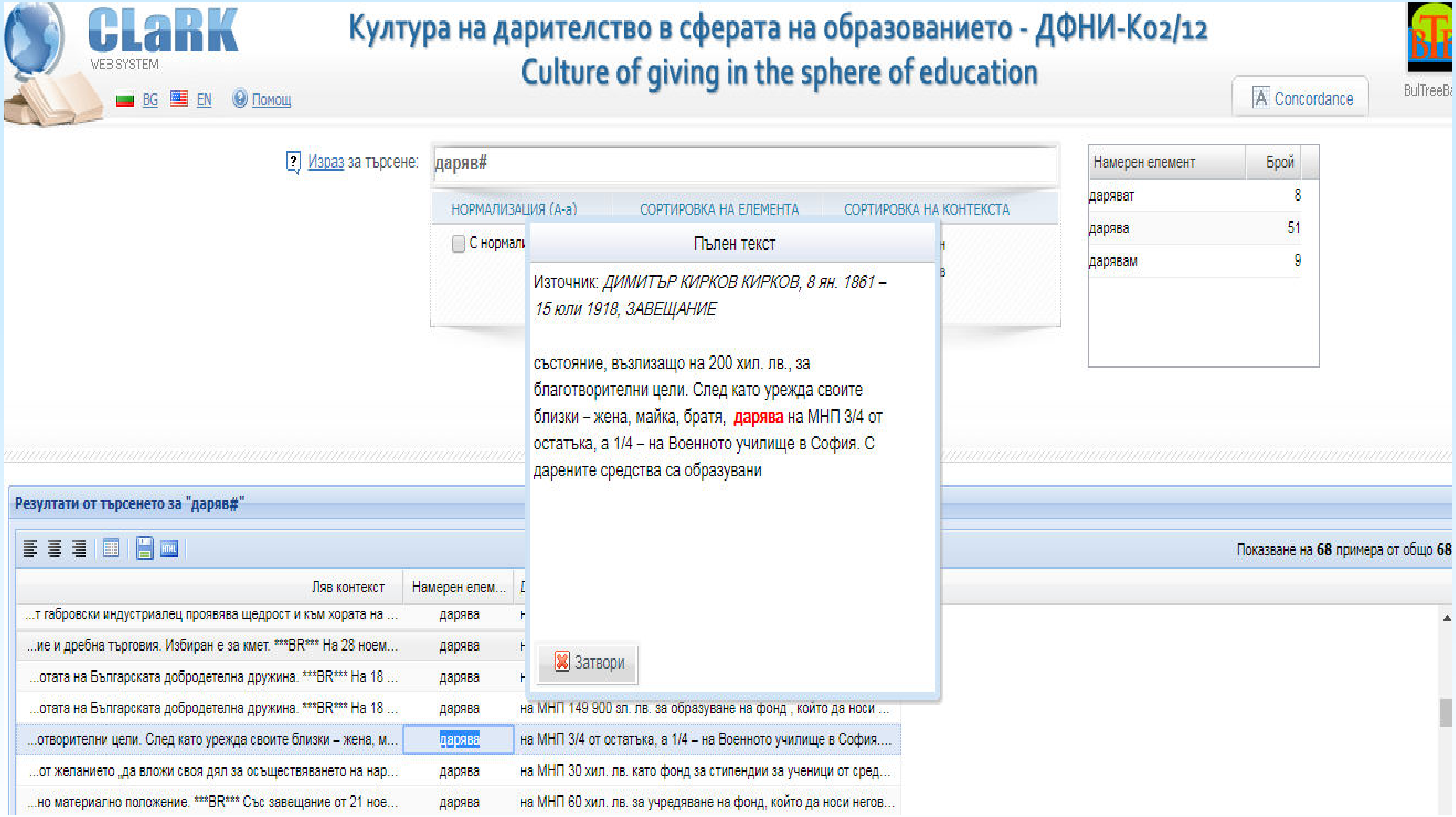
The initial steps include:

- Adaptation of the existing tools,
- The creation of a specialized corpus,
- The creation of a web-based concordance tool, and
- Presenting useful statistics and content analysis over the corpus.

The Specialized Corpus on Giving: CoDar (1)

- Consists of separate documents from the period after the liberation of Bulgaria (from 1878 onward) until the middle of XX century.
- Genre specifics: last will documents; various acts of giving - letters, notarized acts of giving; constitutive documents of charity funds and foundations.
- Web access: http://dar.webclark.org/

Web Concordance



Култура на дарителство в сферата на образованието - ДФНИ-К02/12
Culture of giving in the sphere of education

Израз за търсене: даряв#

Намерен елемент	Брой
даряват	8
дарява	51
дарявам	9

Пълен текст

Източник: *ДИМИТЪР КИРКОВ КИРКОВ, 8 ян. 1861 – 15 юли 1918, ЗАВЕЩАНИЕ*

състояние, възлизащо на 200 хил. лв., за благотворителни цели. След като урежда своите близки – жена, майка, братя, **дарява** на МНП 3/4 от остатъка, а 1/4 – на Военното училище в София. С дарените средства са образувани

Резултати от търсенето за "даряв#"

Показване на **68** примера от общо **68**

Ляв контекст	Намерен елем...	Д
...т габровски индустриалец проявява щедрост и към хората на ...	дарява	Н
...ие и дребна търговия. Избиран е за кмет. ***BR*** На 28 ноем...	дарява	Н
...отата на Българската добродетелна дружина. ***BR*** На 18 ...	дарява	Н
...отата на Българската добродетелна дружина. ***BR*** На 18 ...	дарява	на МНП 149 900 зл. лв. за образуване на фонд , който да носи ...
...отворителни цели. След като урежда своите близки – жена, м...	дарява	на МНП 3/4 от остатъка, а 1/4 – на Военното училище в София....
...от желанието „да вложи своя дял за осъществяването на нар...	дарява	на МНП 30 хил. лв. като фонд за стипендии за ученици от сред...
...но материално положение. ***BR*** Със завещание от 21 ное...	дарява	на МНП 60 хил. лв. за учредяване на фонд, който да носи негов...

The Specialized Corpus on Giving: CoDar (2)

- Three historical periods:
  - *before 1919* (the Bulgarian Renaissance and the end of the First World War) – 49698 word forms;
  - *between 1919 and 1930* (the period of crisis after the First World War) – 46031 word forms, and
  - *after 1930* (the years of stability, the Second World War and the first years after 09.09.1944) – 66373 word forms.
- The texts have been gathered from various libraries and then - scanned and digitized

The Specialized Corpus on Giving: CoDar (3)

- They were represented in an XML format.
- The following types of information were added: metadata, structural and linguistic ones.
  - The *metadata* provides information about: the title of the document and its type (last will, document of giving, etc.), the place and the time of the document emergence; the gender and the social status of the donor/donors.
  - The *structural information* provides the text, divided into paragraphs and sentences.
  - The *linguistic information* provides parts-of-speech, morphosyntactic characteristics and dependency syntactic analysis.

Challenges to NLP

- **For all:** mixture of normalized and authentic texts
- **Tokenizer:** the proper handling of the abbreviations.
- **Morphological analyzer:** rare or archaic words and different orthographical codifications.
- **Lemmatizer:** assigning the word form of a rare word to its lemma
- **NE recognizer:** Person, Location, Organization, Date, Amount
- **Parser:** syntactically different codifications in the contemporary Bulgarian and the texts in previous times

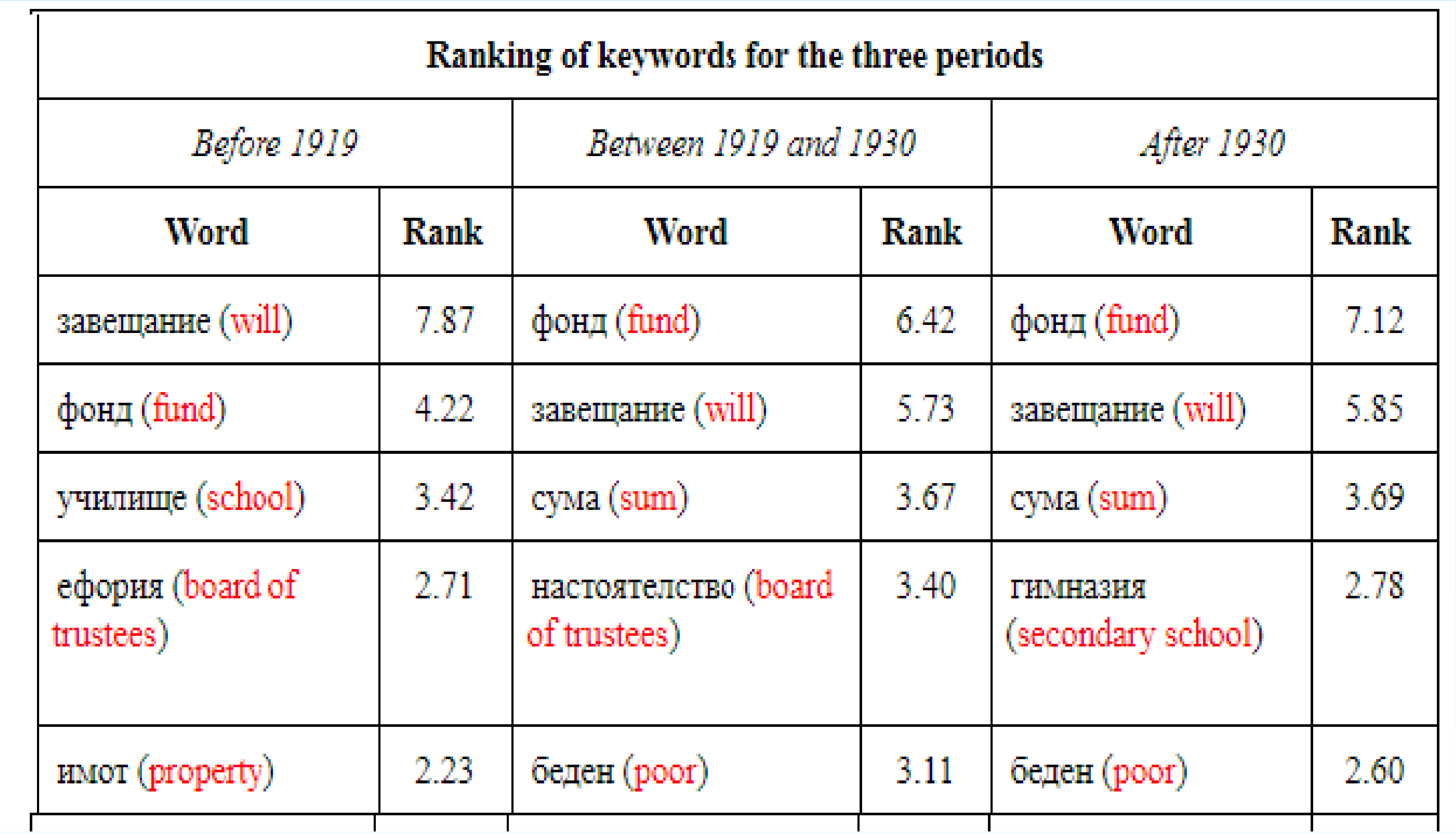
Analyses

- Linguistic Analysis
- Statistical analysis of the keywords that are important for the domain
- Observation of words and phrases in their immediate context (concordance)

Linguistic Analysis: Morpho-tagging

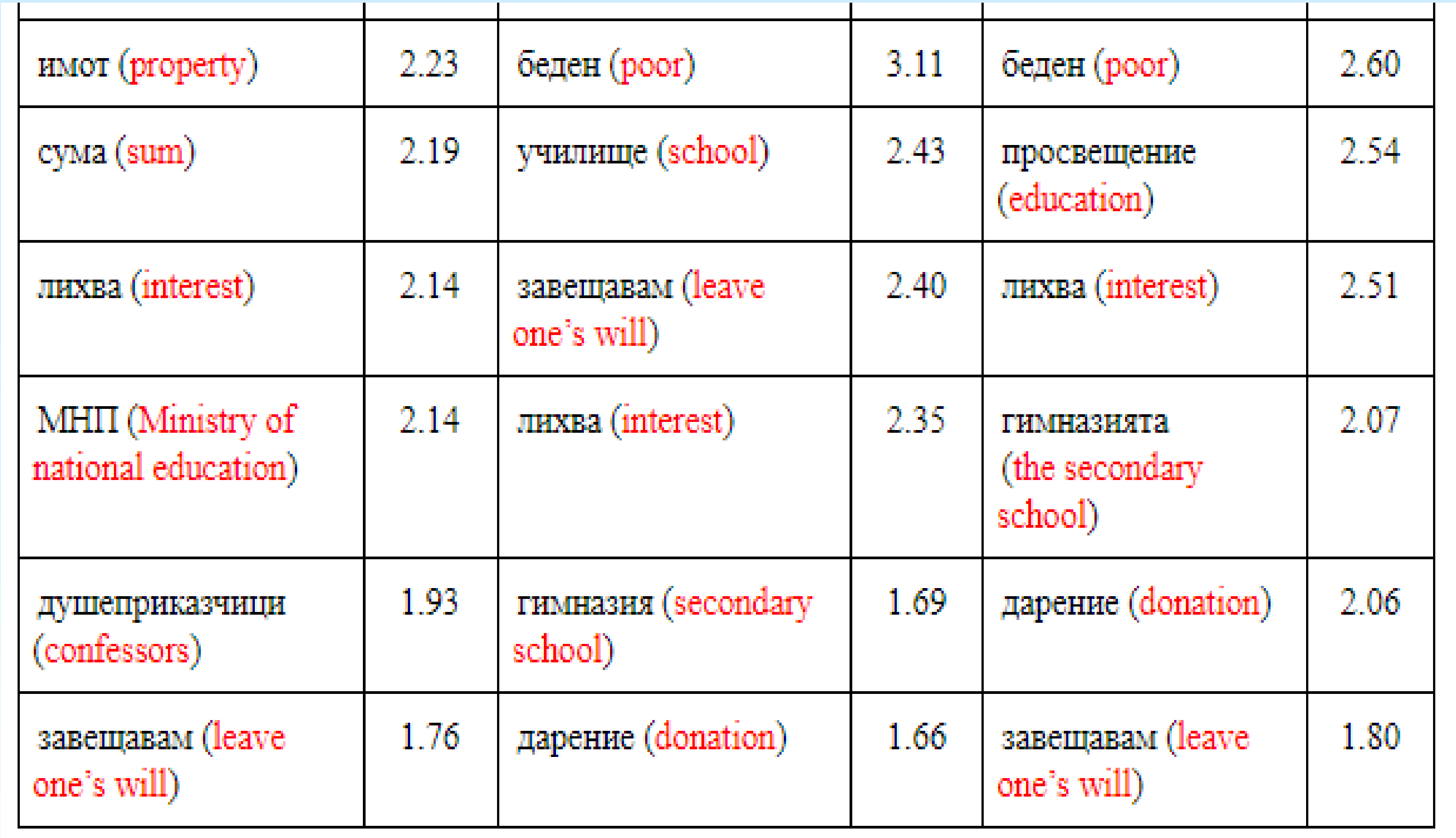
- Qualitative:
  - All periods: dates, names, abbreviations, old case, old orthographic forms: malpropisms – мие (wash) вм. ми е (to me); са (are) вм. се (refl)
  - *Period 1:* same as all periods
  - *Period 2:* plus wrong gender, wrong POS
  - *Period 3:* plus wrong gender, wrong POS, wrong name parts
- Quantitative:
  - *Period 1:* 3.3 % (error rate)
  - *Period 2:* 7 % (error rate)
  - *Period 3:* 12 % (error rate)

The First Ten Most Frequent Words from the Lists with Ranked Keywords



Ranking of keywords for the three periods					
Before 1919		*Between 1919 and 1930*		*After 1930*	
Word	**Rank**	**Word**	**Rank**	**Word**	**Rank**
завещание (will)	7.87	фонд (fund)	6.42	фонд (fund)	7.12
фонд (fund)	4.22	завещание (will)	5.73	завещание (will)	5.85
училище (school)	3.42	сума (sum)	3.67	сума (sum)	3.69
ефория (board of trustees)	2.71	настоятелство (board of trustees)	3.40	гимназия (secondary school)	2.78
имот (property)	2.23	беден (poor)	3.11	беден (poor)	2.60

The First Ten Most Frequent Words from the Lists with Ranked Keywords



имот (property)	2.23	беден (poor)	3.11	беден (poor)	2.60
сума (sum)	2.19	училище (school)	2.43	просвещение (education)	2.54
лихва (interest)	2.14	завещавам (leave one’s will)	2.40	лихва (interest)	2.51
МНП (Ministry of national education)	2.14	лихва (interest)	2.35	гимназията (the secondary school)	2.07
душеприказчици (confessors)	1.93	гимназия (secondary school)	1.69	дарение (donation)	2.06
завещавам (leave one’s will)	1.76	дарение (donation)	1.66	завещавам (leave one’s will)	1.80

Tendencies Analysis

- Terminology change: from *ephoria* to *board*
- *In the period of 1919-1930 and after 1930*: the adjective *poor* comes at 5th position, but lacks among the first 10 before 1919
- *Until 1919* the popular form of charity was through the *property*, but not later

What do Rare Words Say?

- *Before 1919:* the role of the *board members* and *executives*, while in other periods - the *will presenter*
- The *charity related words* have average frequency *in all three periods*
- Rare usage of the verb and noun *wish* in all three periods

Example Use Cases

- Search for information on female donors
- Search for the grant receivers
- Search for the supported causes

Female Donors

- Only 20 results, but after filtering - 10 women
- Extracted names, dates of birth and death from metadata and texts
- Prevalence of donations, not wills
- Frequent co-occurrence: *the will of the donor*

Grant Receivers

- 56 results
- Main content concerns:
  - The conditions, under which the grant can be used, such as the realization of the student in a certain area or in Bulgaria
  - The conditions of grant termination

Supported Causes

- 70 results
- The grant receivers are mainly:
  - The schools and gymnasia
  - Poor children
  - Blind children

Conclusions

- The initial attempts were described on
  - The creation of a corpus on charity activities in 3 periods
  - Initial processing of the corpus
  - Initial statistical and content analysis
- Future plans
  - Cleaning of data
  - Normalizing the archaic words
  - Creation of Linked Open Data datasets interconnected with the existing datasets like DBpedia, GeoNames,
  - Re-training of the NLP pipeline on the specific data