

Министерство науки и высшего образования Российской Федерации  
федеральное государственное автономное образовательное учреждение  
высшего образования  
«Национальный исследовательский университет ИТМО»  
Факультет инфокоммуникационных технологий

**Практическая работа №3**  
***«Многоязычные корпуса и корпуса***  
***второго языка»***  
**по**  
**дисциплине:**  
**«Компьютерная лингвистика»**

**Выполнили:**  
студентки II курса ИКТ  
группы K3242  
Ф.И.О.  
Боброва Мария Иосифовна,  
Шикалова Софья Сергеевна,  
Щечелова Дарья Александровна,  
Третьякова Екатерина Сергеевна

**Проверил:**  
доцент ФТМИ  
*Коцюба Игорь Юрьевич*

## ПРАКТИЧЕСКАЯ РАБОТА 3

**Цель работы:** проанализировать корпуса для одних из самых распространенных языков в мире, создать свой корпус выбранного языка

### Задачи:

1. Найти корпуса для испанского, хинди, арабского, бенгальского, португальского языков
2. Выбрать один понравившийся язык
3. Изучить какие корпуса есть у этого языка
4. Составить свой корпус

### Выполнение работы (Задание 7)

Одни из самых распространенных языков в мире - испанский (450 млн говорящих), хинди (больше 360 млн говорящих), арабский (больше 320 млн говорящих), бенгальский (больше 200 млн говорящих), португальский (ок. 200 млн говорящих).

Воспользовавшись каталогами CLARIN ([www.clarin.eu](http://www.clarin.eu)) и ELRA ([www.elra.info](http://www.elra.info)), поисковыми системами Гугл или Яндекс и Википедией, были найдены данные корпуса для каждого языка:

1. Корпус испанского языка
  - а) Ссылки представлены на сайте Национального корпуса русского языка.  
<https://www.corpusdelespanol.org/> (корпус Марка Дэвиса)  
<https://corp.hum.sdu.dk/cqp.es.html> (поиск корпусов CorpusEye).
  - б) Два корпуса испанского языка, созданные Испанской королевской академией: современный корпус CREA (<https://corpus.rae.es/creanet.html>) и исторический корпус CORDE (<https://corpus.rae.es/cordenet.html>)
2. Корпус хинди  
На сайте Национального корпуса русского языка представлена данная ссылка: <https://www.cfilt.iitb.ac.in/~corpus/hindi/>, однако доступ запрещен
3. Корпус арабского языка  
На сайте Национального корпуса русского языка представлена данная ссылка: <https://arabiccorpus.byu.edu/>
5. Корпус бенгальского языка  
Разговорный корпус, представленный в виде аудиофайлов и отдельный документ с транскрипцией файлов: <https://www.openslr.org/37>
5. Португальский язык  
На сайте Национального корпуса русского языка представлены данные ссылки:  
<https://www.corpusdoportugues.org/> (корпус Марка Дэвиса)  
<https://corp.hum.sdu.dk/cqp.pt.html> (сайт CorpusEye)

Выбранный нами язык для анализа и разработки корпуса – бенгальский.

Первым этапом был поиск датасета на бенгальском языке. Нами был найден архив, состоящий из 1891 аудиозаписи фраз на бенгальском языке на коммерчески направленные темы и tsv файл с транскрибированными аудиозаписями, выглядящий следующим образом:

1	ban_00737_00012222450	এইচআর টেক্সটাইল বাংলাদেশের ভেতরে একাধিক আউটলেটের মাধ্যমে শাড়ি বাচ্চাদের পোশাক
2	ban_00737_00015581920	স্ট্যান্ডার্ড ব্যাংক এ ইসলামী ব্যাংকিং এর সুবিধা রয়েছে
3	ban_00737_00028634754	লাফার্জ সুরমা সিমেন্ট সর্বাধিক ব্যবহৃত সিমেন্ট উৎপাদন করে
4	ban_00737_00035050432	পিপলস ইস্যুরেন্স অব চায়না ছেয়ুটি বছর আগে ব্যবসা চালু করে
5	ban_00737_00068052117	বয়গেস একটি ইন্ডাস্ট্রিয়াল গ্রুপ
6	ban_00737_00107291991	কেয়া ডেভেলপার্স দেশের বিভিন্ন স্থানে স্থাপনা তৈরি করে থাকে
7	ban_00737_00112921837	ভেরাইজিন কমিউনিকেশনস একটি আমেরিকান ব্রডব্যান্ড ও টেলিযোগাযোগ কোম্পানি
8	ban_00737_00120125731	মেটলাইফ বিশ্বের দ্রুত বিসতৃত ইস্যুরেন্স কোম্পানি
9	ban_00737_00120232454	সাইফ পাওয়ারটেক প্রাইভেট কোম্পানি হিসেবে নিবন্ধিত এবং পরিচালিত
10	ban_00737_00142194715	অলিম্পিক ইন্ডাস্ট্রিজ জীবনকে সহজ করে তুলেছে
11	ban_00737_00176672702	বেইজিং অটোমোটিভ গ্রুপ একটি চীনা রাষ্ট্রীয় উদ্যোগ ও নিয়ন্ত্রণকারী কোম্পানি
12	ban_00737_00180288955	পরিবহন খাতে আফতাব অটোমোবাইলস একটি অতি পরিচিত নাম
13	ban_00737_00183428707	চায়না কমিউনিকেশনস কনস্ট্রাকশন চীনের বৃহত্তম বন্দর নির্মাণ ও নকশা এবং ড্রেজিং কোম্পানি
14	ban_00737_00215879121	মেঘনা কনডেন্সড মিল্ক ইন্ডাস্ট্রিজ এর অতি জনপ্রিয় ব্র্যান্ড হলো ফ্রেশ
15	ban_00737_00272549577	মারুবেনি তার কোম্পানির বিশ্বাস হিসেবে সত্যতা উদ্ভাবন ও একাধার করে
16	ban_00737_00295500853	ডাউ কেমিক্যাল আমেরিকান কেমিক্যাল কাউন্সিলের সদস্য
17	ban_00737_00317863452	লিবার্টি মিউচুয়াল ইস্যুরেন্স গ্রুপ একটি আমেরিকান বিমা কোম্পানি
18	ban_00737_00359383339	ফেডারেল ইস্যুরেন্স বাংলাদেশের একটি শীর্ষস্থানীয় বেসরকারী সাধারণ বিমা কোম্পানি
19	ban_00737_00371172031	শাহজালাল ইসলামী ব্যাংক শিক্ষার্থীদের জন্য বৃত্তির করেছে
20	ban_00737_00376983060	জেবিএস একটি ব্রাজিলিয়ান কোম্পানি যারা মাংস প্রক্রিয়াকরণ ব্যবসা করে থাকে
21	ban_00737_00399904268	ঢাকা ব্যাংক ক্রেতার চাহিদার কথা মাথায় রেখে প্রকল্প নির্বাচন ও বাস্তবায়ন করে থাকে
22	ban_00737_00402143113	বেক্সিমকো টেক্সটাইল এর বানিজ্যিক প্রধান শাখা ধানমন্ডিতে অবস্থিত
23	ban_00737_00422977360	বেয়ার নাৎসি যুদ্ধাপরাধে অংশগ্রহণ করার জন্য প্রথম বিশ্বযুদ্ধের পর এবং দ্বিতীয় বিশ্বযুদ্ধের স
24	ban_00737_00460641213	লিভে বাংলাদেশ দীর্ঘ সময় ধরে বাংলাদেশে অপারেশন ও ব্যবসা প্রতিষ্ঠান প্রসার করে আসছে
25	ban_00737_00530142458	বাংলাদেশ শিপিং কর্পোরেশন একটি রাষ্ট্রীয় মালিকানাধীন ও পরিচালিত সরকারি কর্পোরেশন এব
26	ban_00737_00535389845	ও এম ভি গ্রুপ একটি সমন্বিত আন্তর্জাতিক তেল ও গ্যাস কোম্পানি
27	ban_00737_00604373891	সাইনোফার্ম চীনের বৃহত্তম ফার্মাসিউটিকাল কোম্পানি
28	ban_00737_00606350601	প্রিমিয়ার সিমেন্ট মিলস এখন অধিক বিক্রিত সিমেন্ট
29	ban_00737_00636200661	এশিয়া প্যাসিফিক জেনারেল ইস্যুরেন্স গ্রাহকদের সর্বোচ্চ নিরাপত্তা প্রদান করে সর্বনিম্ন খরচ ও
30	ban_00737_00669891346	দেশের অন্য যে কোন বেসরকারি ব্যাংকের থেকে পূর্বালী ব্যাংক এর শাখার সংখ্যা বেশি রয়েছে
31	ban_00737_00741113284	সিএনওএসি গণপ্রজাতন্ত্রী চীনের তৃতীয় বৃহত্তম জাতীয় তেল কোম্পানি

Следующим этапом была подготовка данных для дальнейшего создания корпуса.

Выбранным решением было преобразование tsv файла в txt формат. После преобразования файл был очищен от кодовых номеров аудиозаписей и в качестве разделителей между предложениями был выбран знак “;”.

Файл приобрел следующий вид и был готов к дальнейшей обработке:

এইচআর টেক্সটাইল বাংলাদেশের ভেতরে একাধিক আউটলেটের মাধ্যমে শাড়ি বাচ্চাদের পোশাক মহিলাদের পোশাক এবং অন্যান্য টেক্সটাইল পণ্য উৎপাদন ও বিপণন করে ;

স্ট্যান্ডার্ড ব্যাংক এ ইসলামী ব্যাংকিং এর সুবিধা রয়েছে ;

লাফার্জ সুরমা সিমেন্ট সর্বাধিক ব্যবহৃত সিমেন্ট উৎপাদন করে ;

পিপলস ইন্স্যুরেন্স অব চায়না ছেখ্টি বছরে আগে ব্যবসা চালু করে ;

বয়গেস একটি ইন্ডাস্ট্রিয়াল গ্রুপ ;

কেয়া ডেভেলপারস দেশের বিভিন্ন স্থানে স্থাপনা তৈরি করে থাকে ;

ভেরাইজন্স কমিউনিকেশনস একটি আমেরিকান ব্রডব্যান্ড ও টেলিযোগাযোগ কোম্পানি ;

মেটলাইফ বিশ্বের দ্রুত বিস্তৃত ইন্স্যুরেন্স কোম্পানি ;

সাইফ পাওয়ারটেক প্রাইভেট কোম্পানি হিসেবে নিবন্ধিত এবং পরিচালিত ;

অলিম্পিক ইন্ডাস্ট্রিজ জীবনকে সহজ করে তুলেছে ;

বেথজিং অটোমোটিভ গ্রুপ একটি চীনা রাষ্ট্রীয় উদ্যোগ ও নিয়ন্ত্রণকারী কোম্পানি ;

পরিবহন খাতে আফতাব অটোমোবাইলস একটি অতি পরিচিত নাম ;

চায়না কমিউনিকেশনস কন্সট্রাকশন চীনের বৃহত্তম বন্দর নির্মাণ ও নকশা এবং ড্রেজিং কোম্পানি ;

মেঘনা কনডেলস্ট মিল্ক ইন্ডাস্ট্রিজ এর অতি জনপ্রিয় ব্র্যান্ড হলো ক্রেস ;

মার্কবেনি তার কোম্পানির বিশ্বাস হিসেবে সততা উদ্ভাবন ও ঐক্য ঘারণ করে ;

ডাউ কেমিক্যাল আমেরিকান কেমিক্যাল কার্ডিলের সদস্য ;

লিবাটি মিউচুয়াল ইন্স্যুরেন্স গ্রুপ একটি আমেরিকান বিমা কোম্পানি ;

ফেডারেল ইন্স্যুরেন্স বাংলাদেশের একটি শীর্ষস্থানীয় বেসরকারী সাধারণ বীমা কোম্পানী ;

শাহজালাল ইসলামী ব্যাংক শিক্ষার্থীদের জন্য বৃত্তির করেছে ;

জেবিএস একটি ব্রাজিলিয়ান কোম্পানি যারা মাংস প্রক্রিয়াকরণ ব্যবসা করে থাকে ;

ঢাকা ব্যাংক ক্রেতার চাহিদার কথা মাথায় রেখে প্রকল্প নির্বাচন ও বাস্তবায়ন করে থাকে ;

বেল্লিমকো টেক্সটাইল এর বানিজ্যিক প্রধান শাখা থানমন্ডিতে অবস্থিত ;

বোয়ার নাংসি যুদ্ধাপরবে অংশগ্রহণ করার জন্য প্রথম বিশ্বযুদ্ধের পর এবং দ্বিতীয় বিশ্বযুদ্ধের সময় যুক্তরাষ্ট্রে ব্যবসা হারিয়েছিল ;

লিভ বাংলাদেশ দীর্ঘ সময় ধরে বাংলাদেশে অপারেশন ও ব্যবসা প্রতিষ্ঠান প্রসার করে আসছে ;

বাংলাদেশ শিপিং কর্পোরেশন একটি রাষ্ট্রীয় মালিকানাধীন ও পরিচালিত সরকারি কর্পোরেশন এবং বাংলাদেশের সর্ববৃহৎ জাহাজ মালিক ;

ও এম ডি গ্রুপ একটি সমন্বিত আন্তর্জাতিক তেল ও গ্যাস কোম্পানি ;

সাইনোফার্ম চীনের বৃহত্তম ফার্মসিউটিকাল কোম্পানি ;

প্রিমিয়ার সিমেন্ট মিলস এখন অধিক বিক্রিত সিমেন্ট ;

এশিয়া প্যাসিফিক জেনারেল ইন্স্যুরেন্স গ্রাহকদের সর্বোচ্চ নিরাপত্তা প্রদান করে সর্বনিম্ন খরচ ও দাবির দ্রুত নিষ্পত্তির মাধ্যমে ;

দেশের অন্য যে কোন বেসরকারি ব্যাংকের থেকে পুবাণী ব্যাংক এর শাখার সংখ্যা বেশি রয়েছে ;

সিএনওসি গণপ্রজাতন্ত্রী চীনের তৃতীয় বৃহত্তম জাতীয় তেল কোম্পানি ;

জিবিবি পাওয়ার বেসরকারি ইলেকট্রিসিটি জেনারেশন কোম্পানি ;

সাইনোকাম এর মূল ব্যবসার শক্তি কৃষি কেমিক্যাল রিয়েল এস্টেট এবং আর্থিক সেবায় ;

বিএসআরএম স্টিলস বাংলাদেশের প্রথম আইএসও সনদপ্রাপ্ত ইস্পাত কোম্পানি ;

এমএসএন্ডএডি ইন্স্যুরেন্স একটি জাপানি বীমা কোম্পানি ;

এবিবি বছরজ পরিচিত একটি বছরজাতিক কর্পোরেশন ;

এমআই সিমেন্ট ফ্যাক্টরীর ক্রাউন সিমেন্ট বাংলাদেশের সর্বাধিক রপ্তানিকৃত সিমেন্ট ;

ইউনিলিভার এর পণ্য বৈচিত্র্য এক ও অপ্রতিদ্বন্দ্বী ;

ডয়চে ব্যাংক জার্মান মালিকানাধীন বৃহৎ ব্যাংক ;

বেল্লিমকো নিটিং উচ্চমানের বুনন বস্ত্র উৎপাদন এবং বিক্রি করে ;

বাংলাদেশ সাবমেরিন ক্যাবল কোম্পানি এর লক্ষ্য হলো প্রধান কোম্পানি হিসেবে দেশব্যাপী সাবমেরিন ক্যাবলের টেলিযোগাযোগ সেবা প্রদান করা ;

বাজার মূলধন ও গ্রাহকদের পরিগ্রেহিতে অস্ট্রেলিয়ায় চার বৃহত্তম আর্থিক প্রতিষ্ঠানের একটি ন্যাশনাল অস্ট্রেলিয়া ব্যাংক ;

После этого в качестве инструмента для работы с данными был выбран `bnlp toolkit`<sup>1</sup>. Этот инструмент был специально разработан для работы с бенгальским языком и реализован в качестве библиотеки `ruython`. Нами было выбрано провести исследование какие части речи встречаются чаще в данном датасете. Для реализации данного исследования было решено использовать метод `Pos Tagging`.

Возникшая проблема заключалась в том, что общепринятый список тэгов для английского языка кардинально отличается от тэгов для бенгальского, а `bnlp toolkit` практически не обладает никакой документацией и списка тэгов там не было.

Решение этой проблемы было найдено благодаря статье<sup>2</sup>, в которой представлена таблица тэгов для бенгальского языка.

<sup>1</sup> [Bengali Natural Language Processing\(BNLP\) — bnlp latest documentation](#)

<sup>2</sup> [ICO 2020 Paper 163.pdf](#)

## Сравнение POS Tags списков для английского и бенгальских языков:

### Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential <i>there</i>
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP\$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	<i>to</i>
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP\$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Рисунок 1 - POS Tags для английского языка

**Table 1.** Summary of POS tagsets

11 Tagset	Tagset Count	30 Tagset	Tagset Count
Noun (N)	44425	Common Noun (NC)	30819
		Proper Noun (NP)	7994
		Verbal Noun (NV)	2985
		Spatio-Temporal Noun (NST)	2627
Verb (V)	14292	Main Verb (VM)	12062
		Auxiliary Verb (VAUX)	2230
Pronoun (P)	6409	Pronominals (PPR)	5137
		Reflexive (PRF)	362
		Reciprocal (PRC)	15
		Relative (PRL)	448
		WH Pronoun (PWH)	447
Nominal Modifier (J)	14332	Adjective (JJ)	9377
		Quantifier (JQ)	4955
Demonstrative (D)	2876	Absolute Demonstrative (DAB)	2421
		Relative Demonstrative (DRL)	400
		WH Demonstrative (DWH)	55
Adverb (A)	3965	Adverb of Manner (AMN)	1995
		Adverb of Location (ALC)	1970
Participle (L)	573	Verbal Participle (LV)	72
		Conditional Participle (LC)	501
Post Position (PP)	3989	Post Position (PP)	3989
Particle (C)	6704	Coordinating Particle (CCD)	2899
		Subordinating Particle (CSB)	2051
		Classifier Particle (CCL)	324
		Interjection (CIN)	59
		Others (CX)	1371
Punctuation (PU)	13519	Punctuation (PU)	13519
Residual (R)	4348	Foreign Word (RDF)	1873
		Symbol (RDS)	1968
		Others (RDX)	507

**Рисунок 2 - POS Tags для бенгальского языка**

Расшифровка морфологических единиц для английского и бенгальского языка значительно различаются, например, есть более специфические расшифровки для бенгальского языка (в английском – adverb (RB), в бенгальском – 2 различных наречия (AMN, ALC). То же самое касается большинства частей речи.

Далее нами был написан код на языке Python по созданию корпуса бенгальского языка, который позволяет посчитать процент содержания каждого тэга в тексте. Ввести нужный тэг можно с клавиатуры. В качестве обучающей модели была взята модель для Pos Tagging, предложенная на github разработчиков bnlp toolkit. Данная модель позволяет достигнуть 80% точности.



## **Выводы:**

Малое количество таких тэгов как DWH, LV и PRC в целом характерно для датасетов на бенгальском языке (если полагаться на таблицу из статьи, где подсчитано количество тэгов для 7390 предложений на бенгальском, поэтому их отсутствие в выбранном нами датасете неудивительно), тогда как отсутствие тэгов как RDF, RDS и RDX можно считать спецификой коммерчески направленных предложений на бенгальском языке, а отсутствие тэга RDS легко объясняется тем, что предложения являются транскрибированными аудиозаписями.

## **Источники:**

1. [file:///C:/Users/work11pro1/Downloads/ICO\\_2020\\_Paper\\_163%20\(1\).pdf](file:///C:/Users/work11pro1/Downloads/ICO_2020_Paper_163%20(1).pdf)
2. [https://catalog ldc.upenn.edu/docs/LDC2010T24/Annotation\\_Guidelines\\_for\\_Hindi.pdf](https://catalog ldc.upenn.edu/docs/LDC2010T24/Annotation_Guidelines_for_Hindi.pdf)