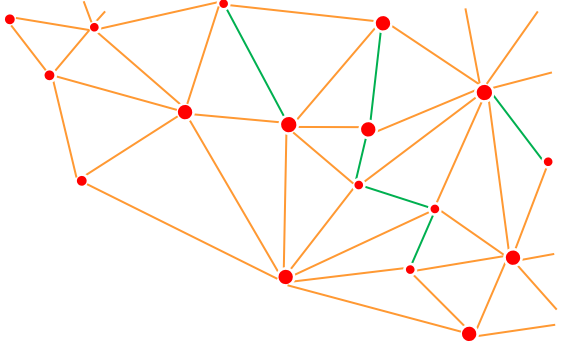


# Small Can Be Beautiful in LLMs for SSH: a Case for Bulgarian – Updated

Nikolay Paev, Kiril Simov, Petya Osenova, Teodor Valchev, Stefan Marinov

IICT – BAS, Bulgaria

BulTreeBank CLaDA BG

# Introduction

Although it is widely accepted that the power of the Large Language Models (LLMs) is in their ability to solve more complex tasks such as

- Question Answering
- Summarization
- Information Retrieval
- Chatbots and many more

we aim at creating a set of LLM-based models for the **basic NLP tasks in Bulgarian**

*"Small" – LM with Transformer-based architecture under 10B*

# Basic NLP tasks

- **Tokenization**
- **Part-of-speech tagging**
- **Lemmatization**
- **Parsing** (constituent or **dependency** syntax)
- **Named Entities Recognition**
- **Named Entities Linking**
- **Word Sense Disambiguation**
- **Event recognition**
- *Coreference Resolution*
- Textual Entailment
- Sentiment Analysis
- Topic modeling

Search engines

Knowledge extraction

# Datasets for Bulgarian

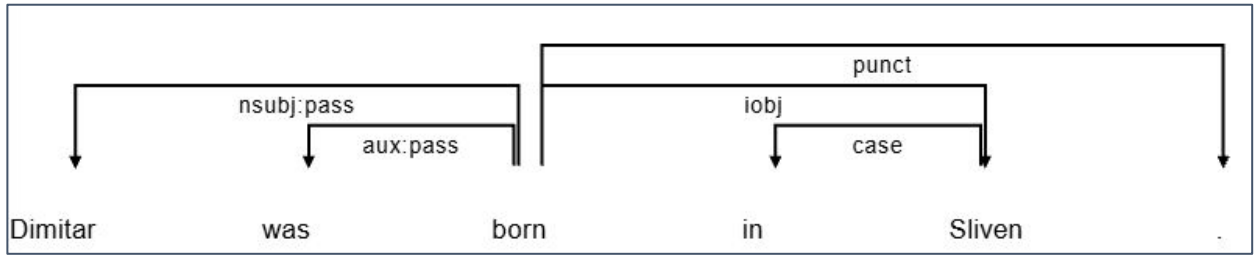
Each of the basic tasks requires to be fine-tuned on a set of appropriate data. For Bulgarian we rely on datasets that are freely available.

**BulTreeBank** (Simov et al., 2002):

- The original constituent variant contained 250 K tokens.
- We extended it on the morphological level with 2.5 M tokens

The corpus is annotated on several layers:

- Morphological characteristics
- Basic word form (lemmas)
- Syntactic structures
- Name Entities



# Datasets for Bulgarian

**Balto-Slavic Corpus (Bulgarian part)** (Piskorski et al., 2021):

- For the shared task the corpus has been annotated with Named Entities (NEs).

On this available corpus, we:

- mapped the NEs to Wikipedia URLs
- processed the texts on the morphological level
- lemmatized all the tokens (400 K)

# Datasets for Bulgarian

**The Bulgarian Event Corpus (BEC)** (Simov et al., 2025):

- BEC comprises 227 documents with 300 K tokens
- The texts are in the area of
  - history
  - biographies
  - ethnography, etc.

The corpus has been annotated with:

- Name Entities with **Links**
- Events (with triggers and roles)
- Coreference chains

**Биография**

**Ранни години**

Роден е на 1 март 1879 г. в Славовица, Пазарджишко. Александър Стамболийски учи в земеделското училище в Садово (1893 – 1895) и завършва Лозаро-винарското училище в Плевен (1895 – 1897). В Плевен е ученик на основателя на Българския земеделски съюз (БЗС) Янко Забунов. През 1899 участва в учредителния конгрес на БЗНС. През следващите години учи философия в Хале и агрономия в Мюнхен, но прекъсва образованието си, поради заболяване от туберкулоза.



Стамболийски като студент в Хале

**Навлизане в политиката**

След завръщането си в България се занимава с политическа дейност. През 1908 година Александър Стамболийски е избран за народен представител от БЗНС. Той се превръща във фактически водач на БЗНС и под негово влияние съюзът е преобразуван от съсловна организация в политическа партия. Народен представител в XIV (1908 – 1911) и XVI-XX (1913 – 1923) обикновено и в V (1911) Велико народно събрание. Под негово въздействие БЗНС провъзгласява лозунга за самостоятелна селска власт, която трябва да защитава „общоселските интереси".



Стамболийски подписва Ньойския договор

# LLM-based NLP Models for Bulgarian Pre-trained Models

- Unsupervised corpora of **29B Bulgarian tokens**
- We sourced them from 3 openly available datasets:
  - *CulturaX (Nguyen et al., 2024)*
  - *Macocu (Bañón et al., 2023)*
  - *HPLT (de Gibert et al., 2024*
- Some other smaller publicly available datasets:
  - Our own data set. Manually cleaned (6B)
  - News sites, Wikipedia, PhD theses, research papers

- The dataset was **deduplicated on document level** with approximated Jaccard Similarity
  *using MinHashLSH from the datasketch package*

# Encoder Models

- For the classification tasks, we developed BERT models of different sizes.
- All of them use a vocabulary of 50 K tokens
- Following the ModernBERT architecture, we also pre-trained a number of models with various sizes

| Model | casing | dim | layers | pars |
|---|---|---|---|---|
| bert-base | both | 768 | 12 | 124M |
| bert-large | both | 1024 | 24 | 355M |
| bert-XL | no | 1024 | 48 | 657M |
| mbert-base | no | 768 | 22 | 149M |
| mbert-large | no | 1024 | 28 | 395M |

# Encoder-Decoder Models

- Not all tasks can be modeled as a classification task.
- Thus, we considered the architecture of **T5 (Text-to-Text Transfer Transformer)** (Raffel et al., 2019a), which combines an encoder and a decoder. We trained different sized T5 models

| Arch | tokenization | dim | layers | pars |
|---|---|---|---|---|
| T5 | subword | 1024 | 12-12 | 403M |
| T5 | character | 1024 | 16-16 | 470M |
| T5 | subword | 1536 | 16-16 | 1.1B |

# LLM-based NLP Models for Bulgarian
# Task specific Fine-tuning

The rows below show results for the following tasks:

- Tokenization and Sentence Segmentation – 0.9940
- Named Entity Recognition – 0.9497
- Part of speech and morphological tagging (XPOS) – 0.9878
- Unlabeled syntax head prediction (UAS) – 0.9571
- Text error tagging – 0.9462
- Word Sense Disambiguation – 0.8244

| **Pre-trained Model** | **Tok+SS** (F1) | **NER** (F1) | **XPOS** (Acc) | **UAS** (Acc) | **Text Err** (F1) | **WSD** (F1) |
|---|---|---|---|---|---|---|
| *bert-web-bg* | 0.9669 | 0.9198 | 0.9787 | 0.9033 | 0.8511 | 0.7756 |
| *bert-web-bg-cased* | 0.9849 | 0.9005 | 0.9810 | 0.9192 | 0.8866 | 0.7507 |
| bert-base | 0.9877 | 0.9289 | 0.9865 | 0.9564 | 0.9214 | 0.7970 |
| bert-base-cased | 0.9927 | **0.9497** | 0.9858 | 0.9488 | 0.9131 | 0.7949 |
| bert-large | 0.9895 | 0.9378 | **0.9878** | 0.9412 | 0.9225 | 0.8155 |
| bert-large-cased | **0.9940** | **0.9497** | 0.9877 | **0.9571** | 0.9245 | 0.8136 |
| modernbert-base | 0.9902 | 0.9174 | 0.9864 | 0.9523 | **0.9462** | 0.8119 |
| modernbert-large | 0.9895 | 0.9206 | 0.9875 | 0.9444 | 0.9423 | **0.8244** |

# Tokenization and Sentence segmentation

- Often tokenization and sentence splitting is the first objective to be completed in the NLP pipe
- Classifies the beginning of the words and the sentences
- We used for training the Bulgarian Event Corpus

| Pre-trained Model | Tok+SS (F1) |
|---|---|
| *bert-web-bg* | 0.9669 |
| *bert-web-bg-cased* | 0.9849 |
| bert-base | 0.9877 |
| bert-base-cased | 0.9927 |
| bert-large | 0.9895 |
| bert-large-cased | **0.9940** |
| modernbert-base | 0.9902 |
| modernbert-large | 0.9895 |

# Part of speech and morphological tagging (XPOS)

- Task of assigning a label corresponding to the grammatical features of the word, grouped in POS tags
- For training we use our extended version of the BulTreeBank *(2.5 M tokens for training)*

| Pre-trained Model | XPOS (Acc) |
|---|---|
| *bert-web-bg* | 0.9787 |
| *bert-web-bg-cased* | 0.9810 |
| bert-base | 0.9865 |
| bert-base-cased | 0.9858 |
| bert-large | **0.9878** |
| bert-large-cased | 0.9877 |
| modernbert-base | 0.9864 |
| modernbert-large | 0.9875 |

# UD parsing for Bulgarian

- Syntax parsing refers to the tree analysis of a given sentence
- We modeled the task in two steps:
  - (1) Annotating the head (parent node) of every word in the sentence (96%)
  - (2) Classifying the syntax relation between each token-head pair (93%)

Our model is working with MWEs in order to improve the results

| Pre-trained Model | UAS (Acc) |
|---|---|
| *bert-web-bg* | 0.9033 |
| *bert-web-bg-cased* | 0.9192 |
| bert-base | 0.9564 |
| bert-base-cased | 0.9488 |
| bert-large | 0.9412 |
| bert-large-cased | **0.9571** |
| modernbert-base | 0.9523 |
| modernbert-large | 0.9444 |

# Named Entity Recognition

- Classical task of assigning labels to word spans that are classified as names
- We use The Bulgarian part of the Balto-Slavic Corpus.
- It contains 11 classes with 4 NER basic categories:
  - Person
  - Location
  - Organization
  - Event
  - Other

| Pre-trained Model | NER (F1) |
|---|---|
| *bert-web-bg* | 0.9198 |
| *bert-web-bg-cased* | 0.9005 |
| bert-base | 0.9289 |
| bert-base-cased | **0.9497** |
| bert-large | 0.9378 |
| bert-large-cased | **0.9497** |
| modernbert-base | 0.9174 |
| modernbert-large | 0.9206 |

# Named Entity Linking

- Named Entity Linking is the task of disambiguating names in a text to a knowledge base
- We labeled the names in the text with their Wikipedia article URL
- We used datasets with Wikipedia URL annotation of named entities:
  - The Bulgarian part of the Balto-Slavic Corpus, the BulTreeBank, and sentences from The Bulgarian Event Corpus (BEC)
  - (53 829 contexts and 7 065 unique named entity urls)

| Position | Score |
|---|---|
| 1 | 0.93 |
| **3** | **0.97** |

# Word Sense Disambiguation

- Word Sense Disambiguation is the task of linking semantically ambiguous words in a text to a thesaurus like WordNet

We modeled the task in two steps:

(1) We use a lexicon and a lemmatizer model to find the candidate senses of the word

(2) We use a cross-encoder model to classify the similarity between the context with the word enclosed in special tokens and the definition of the candidate sense from the Bulgarian BTB-WordNet (Simov and Osenova, 2023)

| Pre-trained Model | WSD (F1) |
|---|---|
| *bert-web-bg* | 0.7756 |
| *bert-web-bg-cased* | 0.7507 |
| bert-base | 0.7970 |
| bert-base-cased | 0.7949 |
| bert-large | 0.8155 |
| bert-large-cased | 0.8136 |
| modernbert-base | 0.8119 |
| modernbert-large | **0.8244** |

# Event Extraction

- Event Extraction is the task of finding and **extracting structured information** from a given text *in the form of events and participants in them*
- Each event is represented as an object and has
  - **type** (birth, death, relocation, etc)
  - **span** – the position in the sentence
  - **list of roles** – each with a **type** and a **span**
- **F1** score of **0.77** in retrieval of event spans
- **Accuracy** of **0.86** in prediction of the type of true positive events

# Lemmatization

- Lemmatization is the task of converting a word to its base form (lemma).
- We fine-tuned a subword level **T5 model**
- Using a train split of our extended version of the **BulTreeBank corpus**
- Achieved a test accuracy score of **0.9946**

*A model is useful in order to cope with Out-Of-Vocabulary word forms, where we can not exploit our inflectional lexicon*

# Text error tagging

- Task of finding spelling, grammatical, and punctuation errors in a given text
- We developed data for this task by noising a text that we consider correct with various rules corresponding to common mistakes, including casing errors
- In this way, we automatically created training dataset of 11M tokens in 200K text chunks

Not a standard NLP task, but very useful in SS&H domain

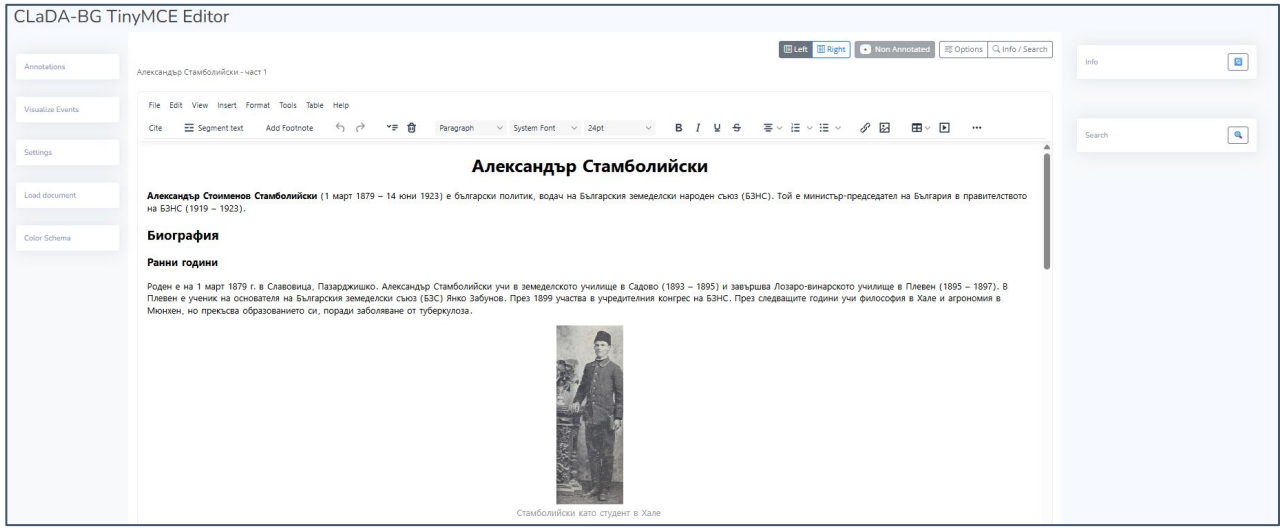
| Pre-trained Model | Text Err (F1) |
|---|---|
| *bert-web-bg* | 0.8511 |
| *bert-web-bg-cased* | 0.8866 |
| bert-base | 0.9214 |
| bert-base-cased | 0.9131 |
| bert-large | 0.9225 |
| bert-large-cased | 0.9245 |
| modernbert-base | **0.9462** |
| modernbert-large | 0.9423 |

# Translation from old spelling

- For this sequence-to-sequence task we use **character level T5**
- In Bulgaria in the 19th and 20th centuries, there were different spelling systems with different letters in comparison to the only one that is used in modern Bulgarian (after 1945)
- *For training, validation and testing data, we used parallel sentences paired from different versions of novels and short stories published in the period of 1910-1944 for which we found also electronic versions published after 1945*
- The best model achieved BLEU score of **0.9532** on the test set

# Visualization of the annotations

- We began developing a user interface to observe the different annotations either in isolation or in combination over documents of interest
- We present a basic user interface implemented as an HTML editor



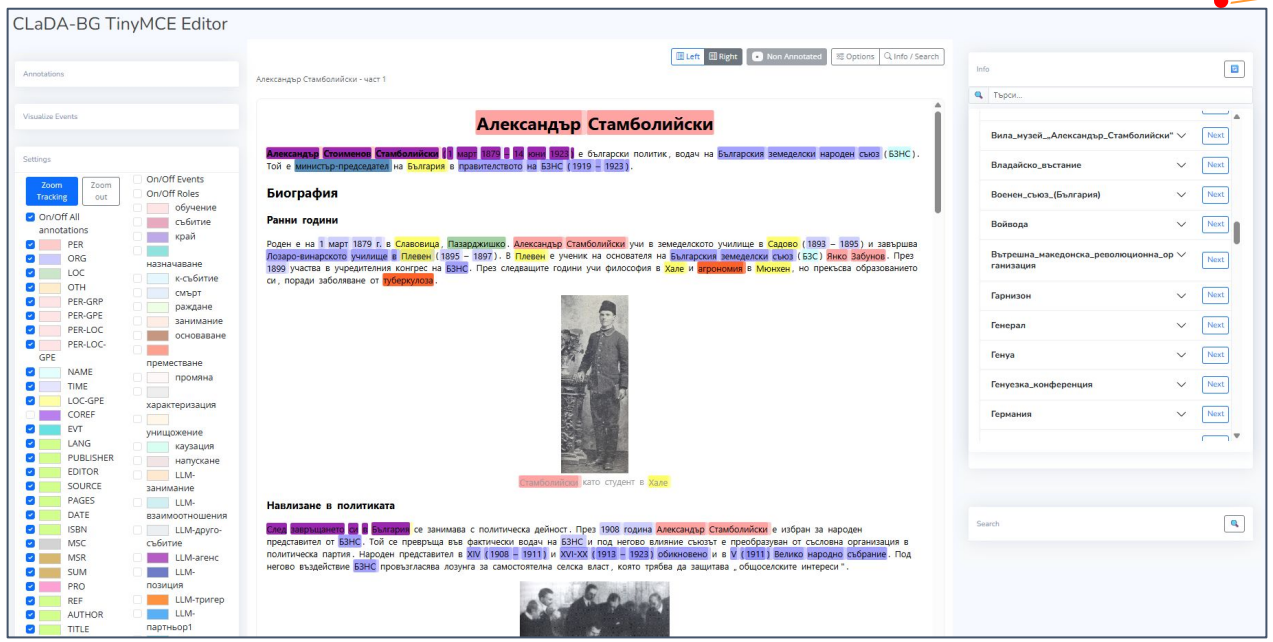
# Visualization of the Annotations

The annotation scheme has three levels of annotation:

- Named entity
- Named entity linking
- Event annotation

Named entity and Named entity linking levels are of type “category”

The Event is represented as a graph between tokens that determines the event in the text




The left part of the screen presents the different categories of Named entities or events. The right part of the screen presents information related to the Named entity links.

# Conclusions

We present a number of LLMs, fine-tuned to basic NLP tasks for Bulgarian. They proved to be useful in supporting research in SS&H:

- POS and lemmatization for searching in large corpora and lexicons
- WSD service is useful for linguists, lexicographers, but also for humanities specialists
- NER and NEL services allow for various representations and thus - specialized access - to one or many documents
- UD Parsing and Event services interact together for more precise event extraction

# Future Work

In our future work, we plan to extend the set of models to new tasks, such as:

- Creation of RAG systems in SS&H
  - 1) models for the vectorization of documents (already implemented)
  - 2) models for the generation of the answers
- Data/Content Enrichment Network for SS&H (and not only)
  - Construct a linked database of documents, knowledge graphs, terminological dictionaries, and others