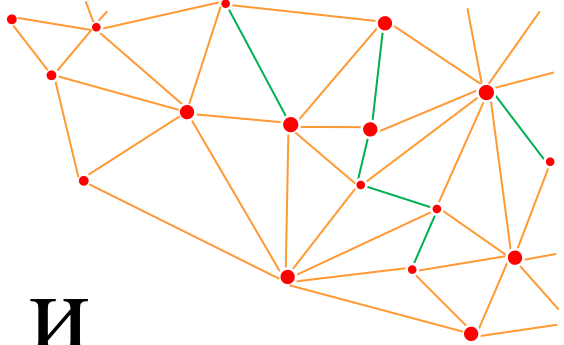
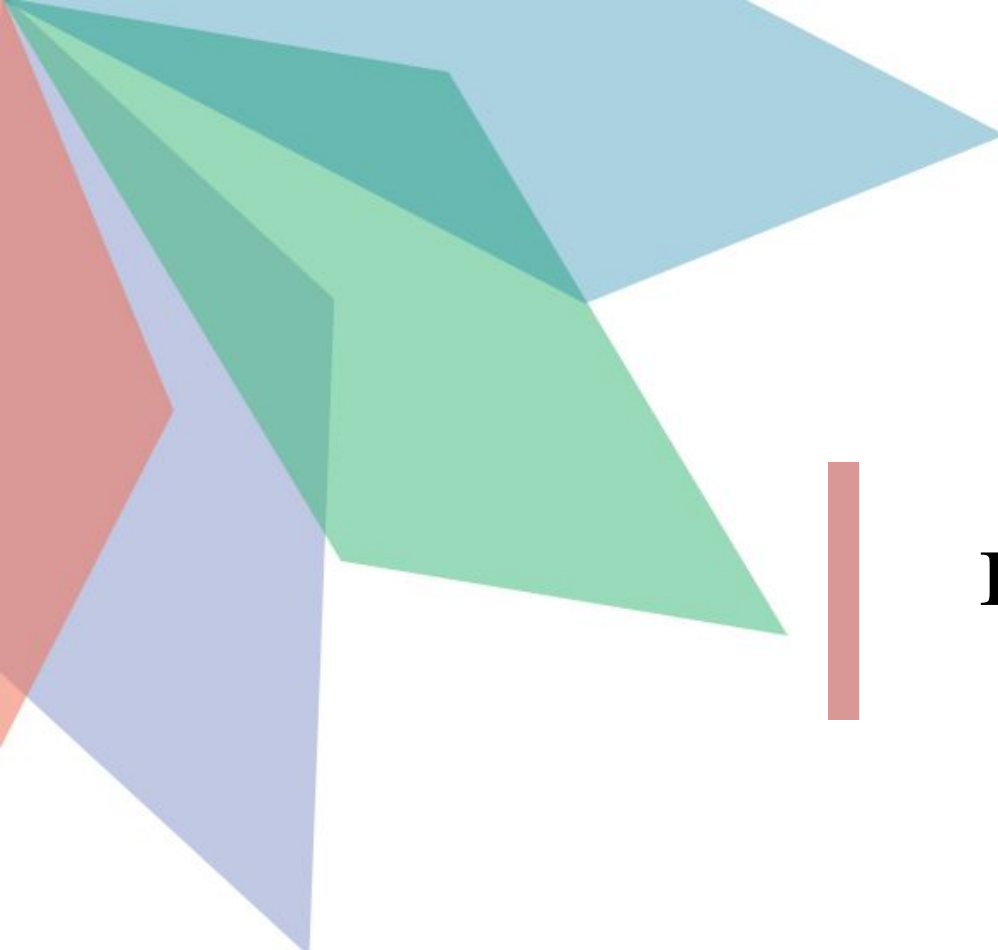


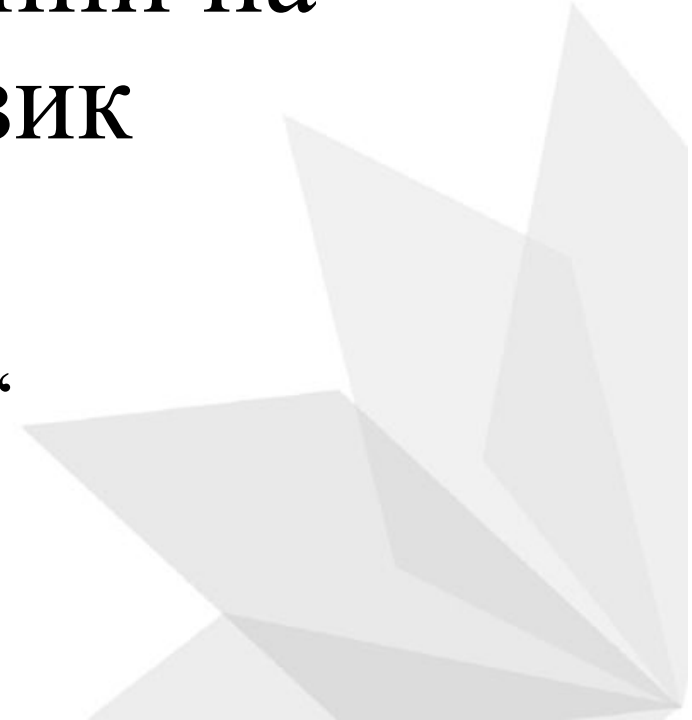


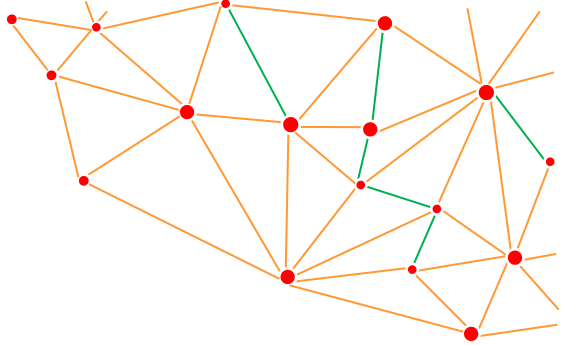
Езикови ресурси и технологии, големи езикови модели за обработка на данни на български език

Петя Осенова

ИИКТ-БАН и СУ „Св. Кл. Охридски“

24.10.2025



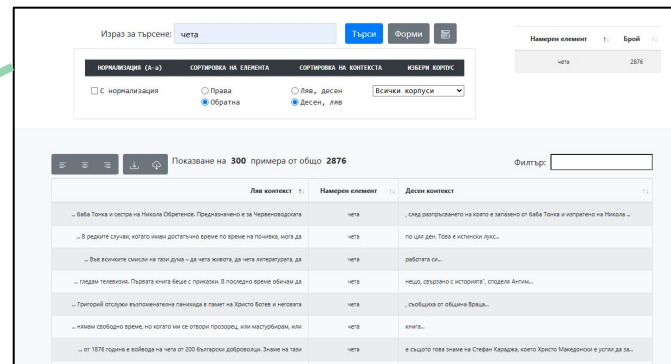


Екип

- Кирил Симов - *координатор*
- Николай Паев - *трениране на Големи езикови модели*
- Живко Ангелов - *концептуализация и бази данни*
- Теодор Вълчев - *потребителски интерфейс и бази данни*
- Стефан Маринов, Кристин Димитрова, Цветелина Александрова, Десислава Нецовска, Александрина Димова, Иоана Попова, Сияна Илиева, Зорница Павлова, Елина Спасова, Калина Шолева, Тихомир Алексиев, Константин Минковски - *анотиране, създаване на ресурси, проверка за качеството*

Езикови ресурси и технологии: роля

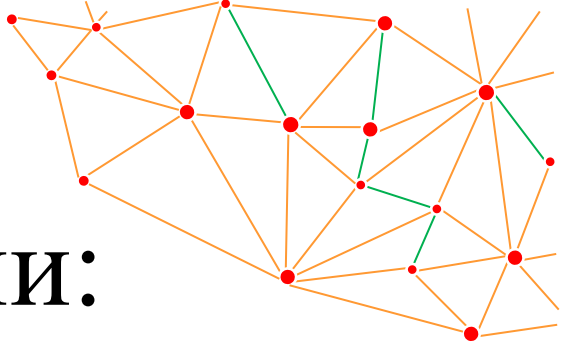
- Подкрепят изследванията в хуманитаристиката чрез създаването на различни езикови ресурси: речници, граматики, корпуси, различни типове упражнения и други
- Моделират знанието в документите, отнасящи се към културното и историческото ни наследство, като ги правят съдържателно достъпни



анотация

ЧЕТА





Езикови ресурси и технологии: съвременни тенденции

- Интегриране на съществуващите ресурси и технологии за подпомагане на различни задачи
- Обогаляване на *Мрежата от знания* за България
- Обучаване на големи езикови модели за определени задачи (AI и GenAI)

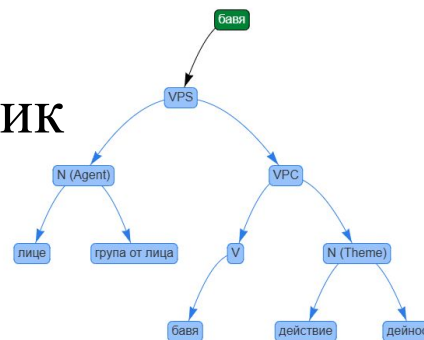
Езикови ресурси и технологии:

Корпуси

- Бултрибанк: синтактично аотиран корпус
- Български морфологично аотиран корпус
- Български корпус със събития
- Корпус със собствени имена
- Български парламентарен корпус (ParlaMint)
- Български интернет корпус CLASSLA
- ПЛАНИРАНИ:
 - Български корпус от периодика преди 1945 г.
 - Български корпус от академични текстове

Езикови ресурси и технологии: речници и лексикални бази

- Мрежа от думи (BTV-WordNet)
- Български валентен речник
- Български фразеологичен речник
- Български речник на обидната лексика



Израз за търсене: партньорка

Търси

Резултат от търсенето

№	Лема	Част от речта
1	партньорка	n

Примери

Нама примери в Мрежата от думи!

Словоизменителен речник: партньорка

партньорка, ~а, ~и, ~ите

Граматична информация

съществително нарицателно, женски род, единствено число, нечленувано

Значения в Мрежата от думи

Лице от женски пол по отношение на човека, с когото е в брак.

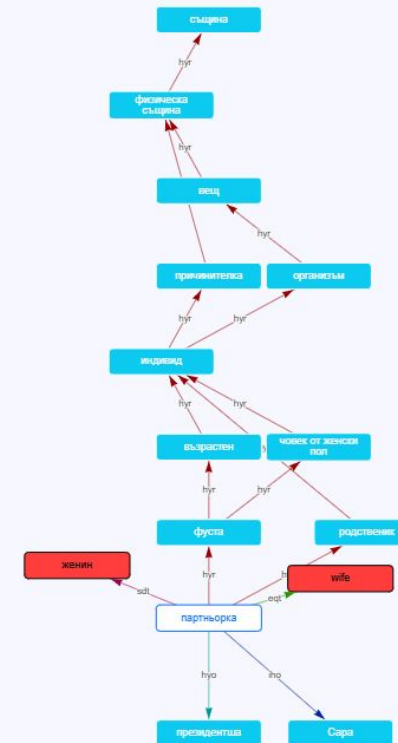
Човек, който участва в някакви действия или дейност с друг.

Български словник за двуезични речници

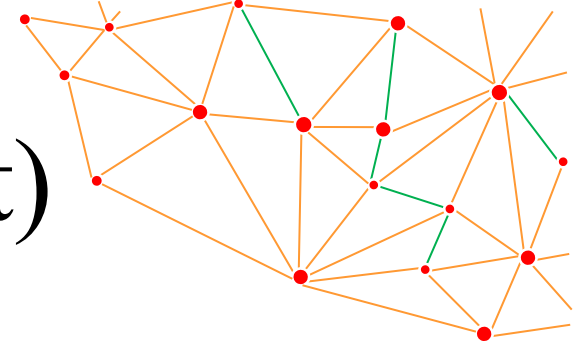
партньор същ. м, [-а, ~и],
партньорка същ. ж, [-а, ~и],
1. Лице, което участва заедно с друг в игра или някаква друга дейност, работа; съучастник, съиграч.
(Търся партньор за шах.)
2. Другар, компаньон. (Млад господин търси подходяща партньорка.)

Словоизменителен състарен речник

партньорка същ. ж., партньорката, партньорки, партньоркиѢ

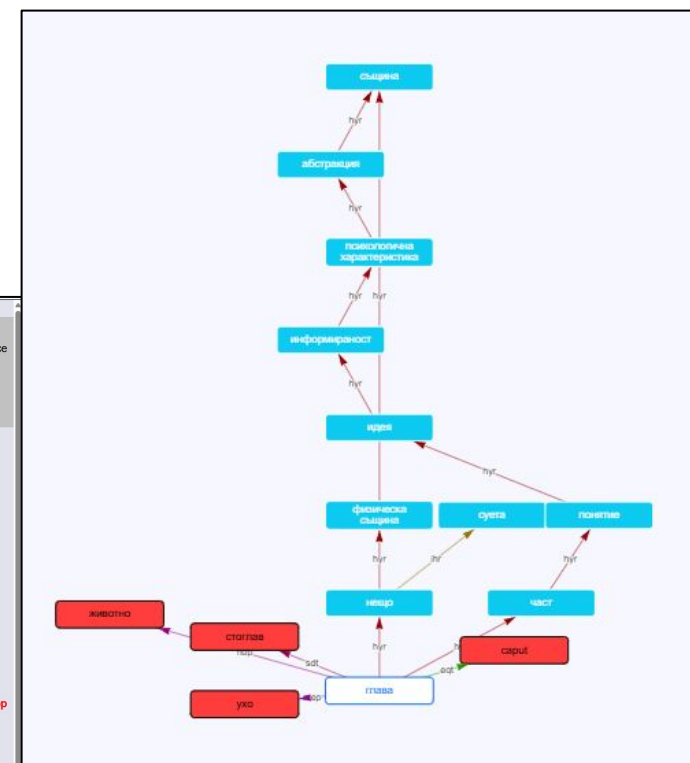


Мрежа от думи (ВТВ-WordNet)

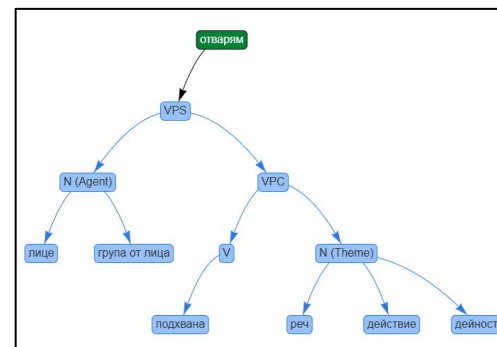
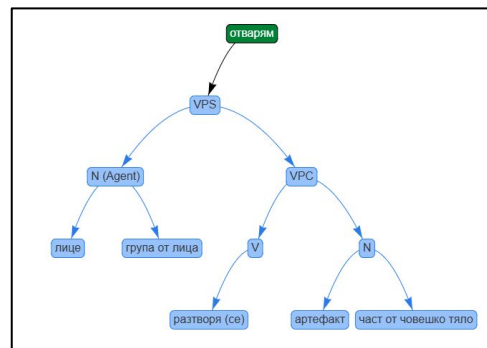
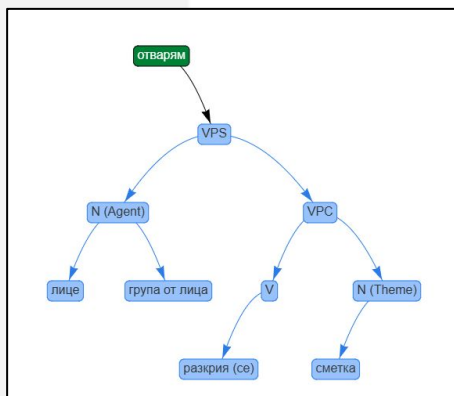
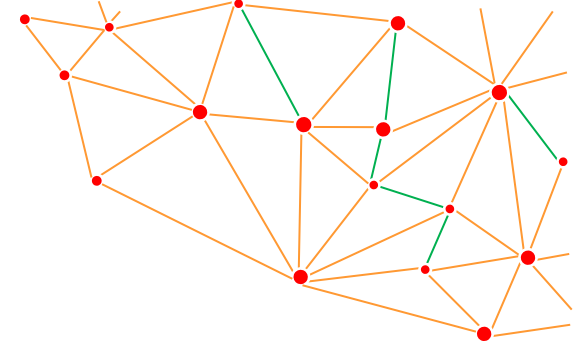


#	Лема	Част на речта	
1	глава	n	(n) китара, кратуна, глава, чутура, тиква, куфалица, главичка, главица Най-горната част на човешкото тяло или предната част на тялото на животно, където се намират мозъкът, устата и повечето сетивни органи. ≈ head Всичко се изглежда от главата ми.
2	главатар	n	
3	главатарка	n	

(n) глава Обособено, относително самостоятелно подразделение на книга, съчинение. ≈ chapter	(n) глава, главичка, главица Разширена, обикновено заоблена крайна част на гвоздей, болт и подобни.
(n) глава, главичка, главица Надебелена заоблена част на някои растения (лук, зеле, цвекло и други).	
(n) глава Лице, което стои начело на семейство, колектив, организация, институция и подобни. ≈ top	dog



Български валентен речник



ВАЛЕНТЕН РЕЧНИК



Намери глагол...

отварям, отвора, откривам, открия, разкривам, разкрива

Wordnet дефиниция: Поставям началото на някаква дейност, работа, предприятие.

Друга дефиниция: Откривам, отварям или поставям началото на някаква дейност.

Категория: verb.change

Скобов анализ: отварям (Agent, Theme | Agent, Theme | Agent) **≈ open up**

Графика



отварям, отвора, отгърщам, отгърна, простирам, простра, разгърщам, разгърна, разпервам, разперя, разварям, разворя

Дефиниция: Разгъвам нещо, опъвам настрана, раздалечавам краищата на нещо.

Категория: verb.contact

Скобов анализ: отварям (Agent) **≈ unfold**

Графика



започвам, започна, пристъпвам, пристъпя, захвана се, захващам се, взема се, вземам се, почвам, почна, залавям се, заловя се, взема, вземам, взимам, захвана, захващам, подхваща, подхващам, начевам, начена, стартiram, отварям, отвора, подкачам, хвана, хващам, зачевам, зачена, подкача, подема, подемам

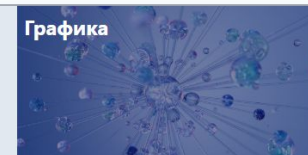
Wordnet дефиниция: Поставям началото на нещо, пристъпвам към някаква работа, дейност, осъществяването, извършването на нещо.

Друга дефиниция: Пристъпвам към нещо; започвам, захващам, подкачам.

Категория: verb.change

Скобов анализ: отварям (Agent, Theme) **≈ begin**

Графика



ID: 1

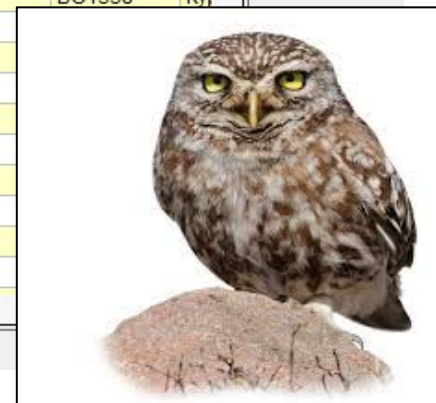
Фразеологизъм [затварям си очите] Дефиниция преносно Премам се, че не забелязвам нещо, съзнателно пренебрегам или не обръщам внимание на нещо, обикновено когато става въпрос за нещо нередно, като избягам да не го виждам или осъзнавам. Употребително преносно Id: 18156 verb perception LEMMA: крия, скривам, скриях, премълчавам, премълчавал, прекривах, прикрих, тай, ставаня, стая, спотайвам, спотая, затаявам, затая, попитаям, попитах, спотичавам, спотичах. Дефиниция буквално Затварям клепачите си, прекратявайки зрителното възприятие за известно време, като може да се направи с цел почивка или за да не виждам нещо конкретно. Употребително буквално Id: 17397 verb contact LEMMA: затварям, затворя DEF: Събирам, съвмам вончия кракцища, части на нещо.	Коментар Част на речта Словоред променлив Класификация Статус ГОТОВ От ревник ЕБТР Глагол затварям Опора Order id 0	Confidence 0.6962158739 Преносно (10) Буквално (8) Реликтьр (1) Ревници (1)
---	---	---

Пример	Източник
▶ Как тя вижда себе си? Харесва ми да съм активна. @@@ Затварям си очите @@@ за проблемите с отношенията с хората. Обичам да с	https://www.pechelimzaedno.net/bg/61427/5974
Отделно – за да оправдавам другите, съм търсела вината в себе си. Не съм искала да остана разочарована от приятели или от хората с	http://skandalno.net/liyana-az-s-mariyan-absurd-intervyuu
В същото време, ние, вестникарите, не бягаме от отговорност. Добре знаем, че думите убиват като куршуми, особено ако уцелват невини	http://www.videlinabg.com/tempDoc/columnLeft/historyS
Австралиеца не е никак пристрастен, чувството на обективност у него е добре развито. Той не обича да @@@ си затваря очите @@@ г	П. Вежинов, ДМ, 50; https://ibl.bas.bg/rbe/lang/bg/zatva
– Има работи, пред които не бива да @@@ си затваряме очите @@@ повече! – избухна Найденов.	В. Геноевска, СГ, 288; https://ibl.bas.bg/rbe/lang/bg/zatva
Асен като ръководител на Беслеменовата фабрика @@@ си затваряше очите @@@ за много работи.	Д. Кисьов, Щ, 475; https://ibl.bas.bg/rbe/lang/bg/zatva
Та що се отнася до нервността, стигнах до простишкия извод, за себе си да не @@@ затварям очите си @@@ за причините, които ме из	http://www.beinsadouno.com/board/forums/topic/11208
Майка ми ме възпита да бъда гражданин на света и да не @@@ затварям очите си @@@ пред суровата реалност.	https://www.ells.bg/a/megan-markul-shte-zasenchli-ke
Пак казвам - Борисов нищо не е. @@@ Затварях си очите @@@ за останалите акули. Които ви уверявах какъв рай е България, а в с	https://www.dnes.bg/politika/2020/05/31/borisov-smenila
Оправдание имат само тези, които не са живяли първите десетилетия в България. Трябвало много силно да си @@@ си затварял очите @	https://glasove.com/categories/novini/news/nominirana

Български речник на обидната лексика

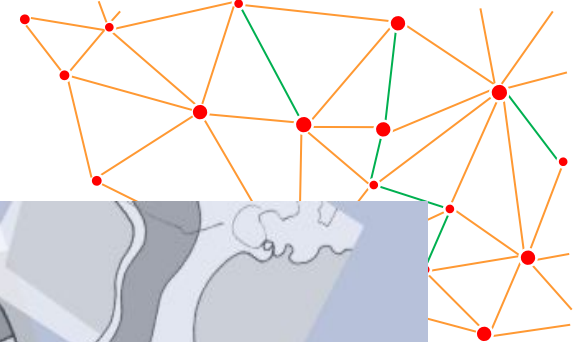
PHRASE	DEF	Is it phrase?	POS	OTHER_PHR	STATUS	BG_ID	OL
идолослужение	Остар. Книж. Служене на идоли, кумири; кумирослужение.	F	noun		Готово	BG1433	Ид
инвалид	Id: 12472 noun.person LEMMA: инвалид DEF: Човек, изгубил изцяло или о	F	noun		Готово	BG1450	Ин
Коза ностра	Сицилианската престъпна организация, мафията	T	noun		Готово	BG1677	Ко
козел	Id: 156469 noun.person LEMMA: козел DEF: Човек с дълга остра брада.	F	noun		Готово	BG1290	Ко
коза	Обидно определение за жена с непривлекателна външност. Особено и	F	noun		Готово	BG680	Ко
колиба	Малка, паянтова къща, обикн. от една стая; къщурка.	F	noun		Готово	BG1409	Ко
конкубина	Прен. В древността е жена, живееща с мъж в открита, нескривана връзка	F	noun		Готово	BG949	Ко
крава	Прен. Разг. Пренебр. Едра и/или глупава жена.	F	noun		Готово	BG484	Кр
крадец	Id: 12958 noun.person LEMMA: джебчия, грабител, крадец, обирджия, тат	F	noun		Готово	BG1325	Кр
красноречие	Id: 141149 noun.communication LEMMA: реторика, риторика, красноречие	F	noun		Готово	BG1770	Кр
кретенизъм	Прен. Глупост, некадърност.	F	noun		Готово	BG866	Кр
крокодил	Нарицателно за грозна жена. И особено за такава, която е с големи и из	F	noun		Готово	BG879	Кр
кукумявка	Id: 143178 noun.person LEMMA: грозник, грозница, плашило, кукумявка D	F	noun		Готово	BG2432	Ку
курва	Id: 11921 noun.person LEMMA: кучка, курва, блудница, развратница, разв	F	noun		Готово	BG1338	Ку
куртизанка	Жена с леко поведение.	F	noun		Готово		
кърлеж	Прен. Безмилостен, жесток господар; мъчител, угнетител, експлоататор.	F	noun		Готово		
леност	Id: 150663 noun.state LEMMA: мързел, леност, безделие, лентяйство, мър	F	noun		Готово		
лентяйство	Id: 150663 noun.state LEMMA: мързел, леност, безделие, лентяйство, мър	F	noun		Готово		
лешояд	Прен. За алчен за материални облаги и безскрупулен човек.	F	noun		Готово		
либертинизъм	Философско течение ограничено в литературните и някои други интелек	F	noun		Готово		
лихвар	Id: 142314 noun.person LEMMA: лихвар, лихварин, лихварка DEF: Човек, к	T	noun		Готово		
лъжа	Id: 8537 noun.cognition LEMMA: измама, лъжа, номер, илюзия, шмекерия	F	noun		Готово		
магаре	Id: 156625 noun.person LEMMA: магаре DEF: Упорит, лош или изобщо ом	F	noun		Готово		

Намерени са 1366 обекта. # INCLUSIVE: 1586 # CONSERVATIVE: 711





Игра на значения :)



Вид тъкан от преплетени и завързани нишки, между които се образуват отвори.

Съвкупност от хардуерни компоненти и компютри, свързани помежду си чрез преносна среда, която позволява обмен на информация помежду им.

Система от взаимно свързани хипертекстови документи, достъпни посредством протокол за прехвърляне на хипертекст.

Тлъста обвивка около червата на животно.



В рамките на КОСТ акцията UniDive: български съответствия на 60 идиома с генерирани изображения

Идиом *на седмото небе съм*: 1. идиоматично, 2. близко до идиоматичното, 3. близко то буквалното, 4. буквалното, 5. дистрактор



1



2



3



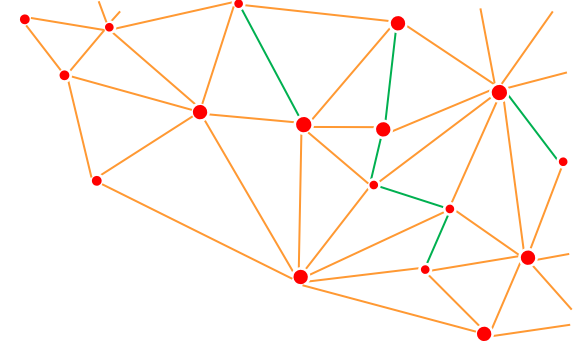
4



5



Езикови ресурси и технологии: равнища на автоматичния анализ

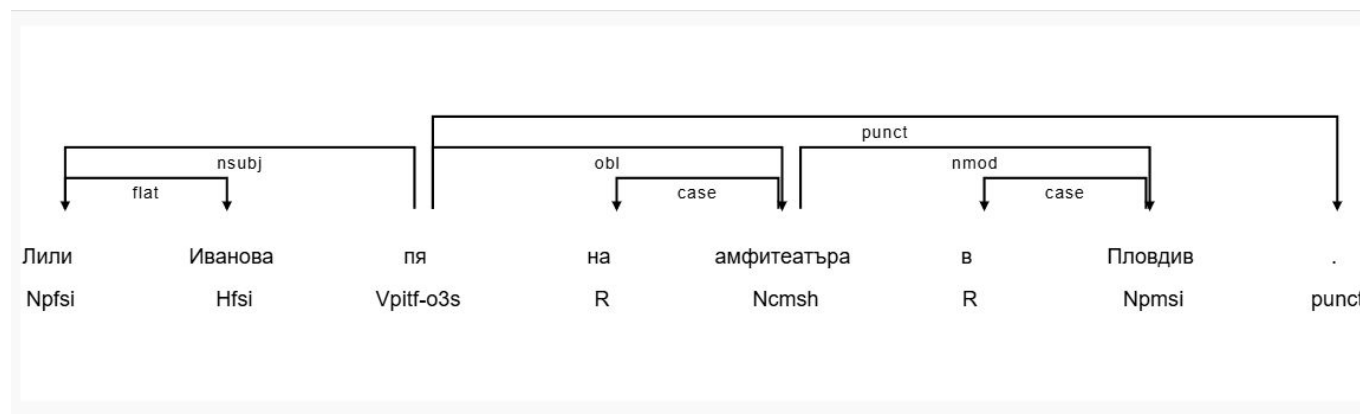


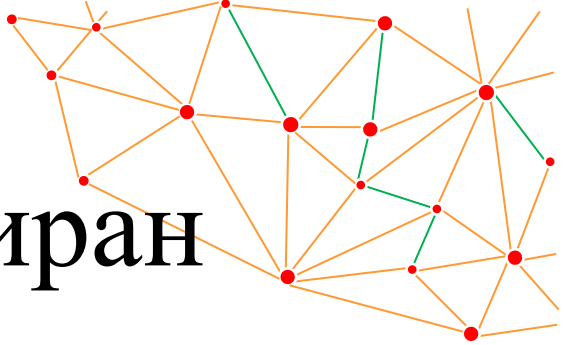
- Сегментация на текста на думи, пунктуация, символи, изречения
- Морфологичен анализ на граматичните характеристики на думите
- Лематизация (намиране на основната форма на думата)
- Синтактичен анализ (конституентен и депендентен)
- Определяне на лексикалните значения на думите
- Разпознаване и класифициране на имена
- Разпознаване на събития и извличане на факти

Синтактичен корпус Бултрибанк

- Корпусът е аотиран с:
 - Граматически характеристики
 - Основна форма на думите (леми)
 - Синтактични структури
 - Собствени имена

- Покритие:
 - 156 000 думи





Български морфологично аотиран корпус

- Корпусът е аотиран с:
 - Граматически характеристики
 - Основна форма на думите (леми)
- Покритие:
 - 2 064 063 думи (без пунктуацията)

Български корпус със събития

- Корпусът е аотиран със:

- Собствени имена
- Събития
- Кореференции

- Покритие:

- 227 документа
- 291 196 думи

Хаджи Димитър



Родът на Хаджи Димитър

Димитър Николов Асенов, по-известен като **Хаджи Димитър**, е един от най-значимите български войводи. Умира едва 28-годишен след тежко сражение на четата му с османски части. По повод на смъртта му **Христо Ботев** написва едноименната поема.

Връх Хаджи Димитър на Антарктическия полуостров е именуван на революционера.

Биография

Хаджи Димитър е роден в **Сливен** в семейството на търговеца **Никола и Маринка Асенови**. Когато е на две години, той и семейството му отиват на поклонение в **Йерусалим**, и оттогава **Димитър** е хаджия.

Хайдушка дейност

През 1862 година **Хаджи Димитър** излиза в Балкана с чета и цяло лято броди из планината. През пролетта на 1864 г. се включва в четата на **Стоян войвода** от **Сливен** като знаменосец. Четата е от 12 души и има за задача да убие търновския гръцки владика **Преда** преди да влязат в **Търново** четата се разпада и четниците се отделят от войводата. Командването се поема от **Хаджи Димитър**, който ги повежда към **Сливенския Балкан**.

раждане

роден

тригер

място

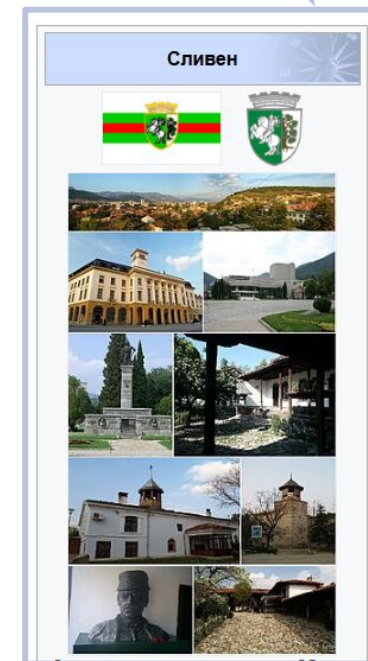
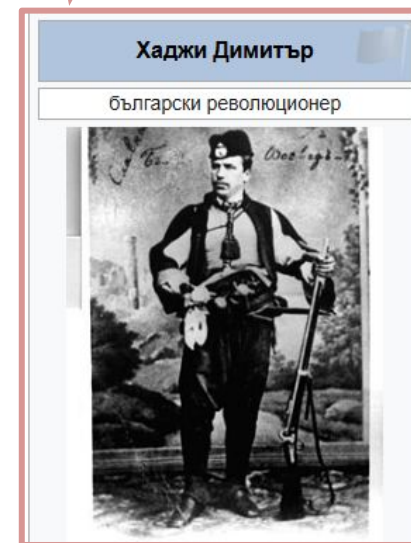
родители

Хаджи Димитър е роден в **Сливен** в семейството на търговеца **Никола и Маринка Асенови**.

Корпус със собствени имена

- Корпусът е аотиран с:
 - Връзки към Уикипедия
 - Морфологична информация
 - Основна форма (лема)
- Покритие:
 - 365 542 думи

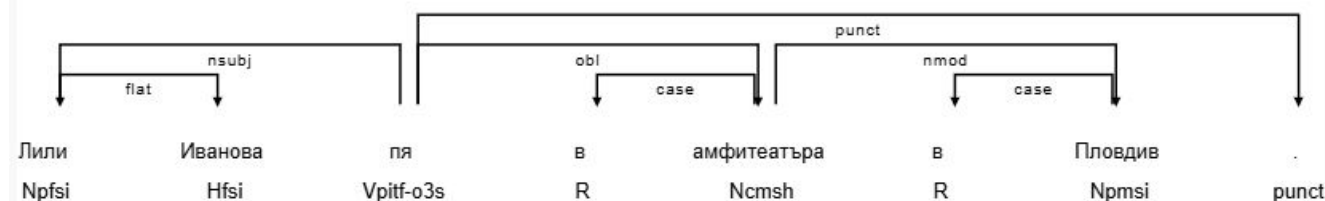
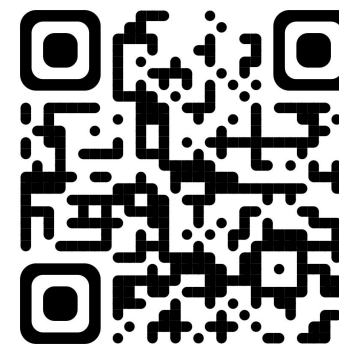
Хаджи Димитър е роден в Сливен в семейството на търговеца Никола Асенов и жена му Маринка.



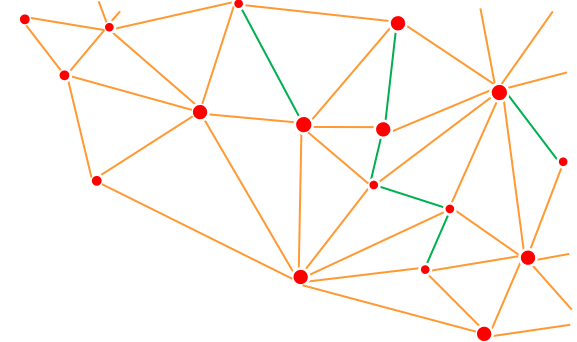
Големи езикови модели: синтактичен анализ с модела BERT

1. Лили Иванова пея в амфитеатъра в Пловдив.

ID	Form	Lemma	UPOS	XPOS	Feats	Head Index	Dep Rel	NER
1	Лили	лили	PROPN	Npfsi	Definite=Ind ...	3	nsubj	B-PER
2	Иванова	иванов	PROPN	Hfsi	Definite=Ind ...	1	flat	I-PER
3	пея	пея	VERB	Vpitr-o3s	Aspect=Imp ...	0	root	O
4	в	в	ADP	R	_	5	case	O
5	амфитеатъра	амфитеатър	NOUN	Ncmsh	Definite=Def ...	3	obl	O
6	в	в	ADP	R	_	7	case	O
7	Пловдив	пловдив	PROPN	Npmsi	Definite=Ind ...	5	nmod	B-LOC
8	.	.	PUNCT	punct	_	2	punct	O

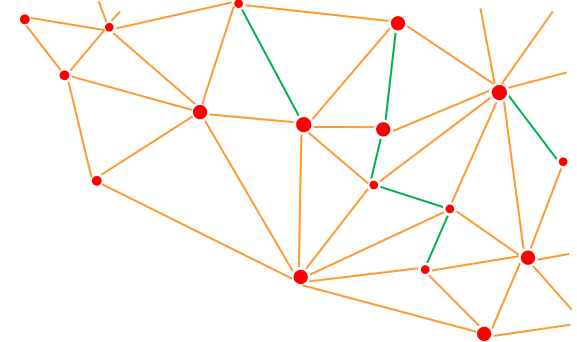


Обучени езикови модели



- Трансформър модели, базирани на архитектурите:
 - *BERT*
 - *ModernBERT*
 - *T5*
- Обучени са върху български текстове в размер на:
 - 29 милиарда думи
- Достъпни чрез платформата *Hugging Face*
 - <https://huggingface.co/AlaLT-IICT>

Сегментация на думи и разделяне на изречения



- Първа стъпка при обработката на текст
- Разпознаване на началото на думи и изречения
- Обучен е BERT (енкодер) модел

Лили|Иванова|пя|в|амфитеатъра|в|Пловдив|.

Морфосинтактична анотация

- Определяне на частите на речта и морфосинтактичните характеристики на думите в текст
- Обучен е BERT (енкодер) модел върху Българския морфологично аотиран корпус
- Точността е 99 %

1. Бела бели бели картофи в Бели Искър и прави бели.

ID	Form	Lemma	UPOS	XPOS
1	Бела	бела	PROPN	Npfsi
2	бели	бял	VERB	Vpitf-r3s
3	бели	бял	ADJ	A-pi
4	картофи	картоф	NOUN	Ncmpi
5	в	в	ADP	R
6	Бели	бял	ADJ	Ams-e
7	Искър	искър	PROPN	Npmsi
8	и	и	CCONJ	Cr
9	прави	прав	VERB	Vpitf-r3s
10	бели	бял	NOUN	Ncfpi
11	.	.	PUNCT	punct

съществително

глагол

прилагателно

съществително

предлог

прилагателно

съществително

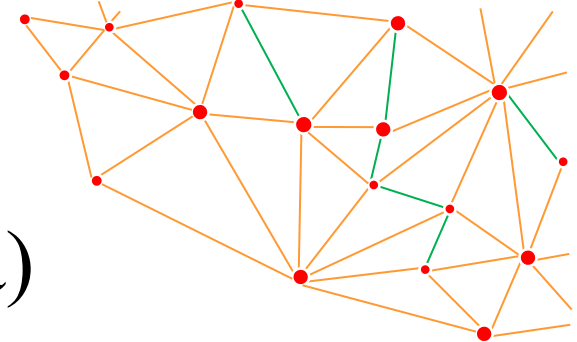
сюз

глагол

съществително

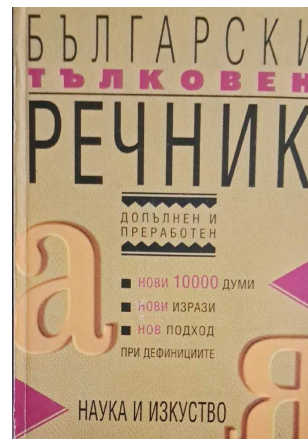
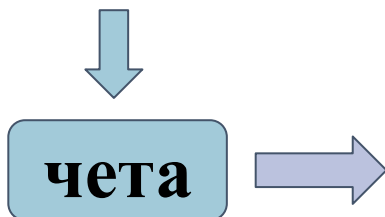
пунктуация

Лематизация (определяне на основната/речниковата форма на думата)



- Обучен е T5 (енкодер-декодер) генеративен модел
- Необходима първа стъпка за разпознаване на думи и свързването им с речници и други ресурси
- Точността е 99 %

Командването се поема от Хаджи Димитър, който повежда **четата** към Сливенския Балкан.



чета същ, ж, [~ата, ~и]

1. Група, отред, дружина въоръжени хора, участници в освободителните борби. (*Четата на Ботев.*)
2. Група лица за обща работа, обикн. поделение на отред. (*Бригадирска чета.*)
3. *остар*, Гимнастическа или друга физкултурна група.

Лематизация (определяне на основната/речниковата форма на думата)

- Обучен е T5 (енкодер-декодер) генеративен модел
- Необходима първа стъпка за разпознаване на думи и свързването им с речници и други ресурси
- Точността е 99 %

Командването се поема от Хаджи Димитър, който повежда **четата** към Сливенския Балкан.

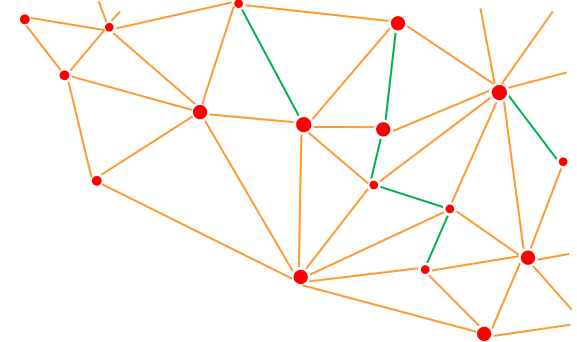


чета

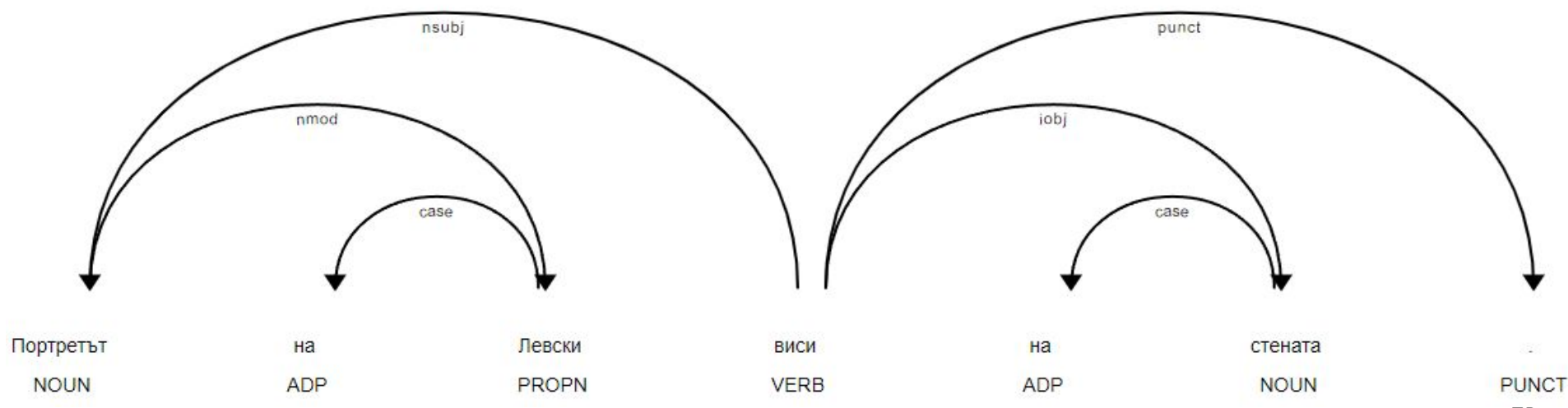


The screenshot shows the Clank web interface. At the top, there's a navigation bar with 'Clank', 'Думата в контекст', 'Всичко за думата', and 'БТБ Мрежа от думи'. The main content area is titled 'Интегрирана система за корпуси и речници'. It features a search bar with 'чета' entered. Below the search bar, there's a table titled 'Резултат от търсенето' with columns '№', 'Лекс', and 'Част от речта'. The table lists three results for 'чета'. To the right, there's a section 'Синонимизирания речник: чета' showing 'чета, -ата, -и, -ите'. Below that, 'Граматична информация' shows 'существително нарицателно, женски род, единствено число, нечленувано'. At the bottom, there's a 'Примери' section with text snippets. On the right side of the interface, there's a section 'Значения в Мрежата от думи' with two entries: 'Група, отред, дружина въоръжени хора, участници в освободителните борби.' and 'Група лица за обща работа, действащи съгласувано и посветени на общата цел.'

Депендентен синтактичен анализ



- Разпознаване на синтактичната структура на изречение
- Обучен BERT модел върху корпуса Бултрибанк
- Анализ в 2 стъпки:
 - Разпознаване на опората на всяка фраза – с 96 % точност
 - Последващо разпознаване на синтактичната релация – с 93% точност



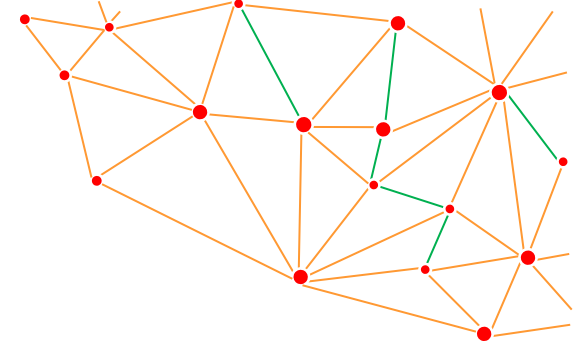
Разпознаване на собствени имена

- Разпознаване на имена и категориите им в текст
- Обучен е BERT (енкодер) модел
- Индексиране на документи спрямо имената в тях
- Macro F1 е 96 %

word	Портретът	на	Левски	виси	на	стената	.
UPOS	NOUN	ADP	PROPN	VERB	ADP	NOUN	PUNCT
NER	O	O	B-PER	O	O	O	O

word	Левски	отново	се	класира	за	Лига	Европа	.
UPOS	PROPN	ADV	PRON	VERB	ADP	NOUN	PROPN	PUNCT
NER	B-ORG	O	O	O	O	B-OTH	I-OTH	O

word	Влакът	пристига	в	град	Левски	привечер	.
UPOS	NOUN	VERB	ADP	NOUN	PROPN	NOUN	PUNCT
NER	O	O	O	O	B-LOC	O	O

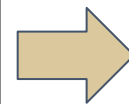


Осъвременяване на стар правопис

- Осъвременяване на текст със стария правопис (преди реформата от 1945 г.)
- Използване на T5 (енкодер-декодер) модел на ниво букви
- BLUE мярка: 96 %

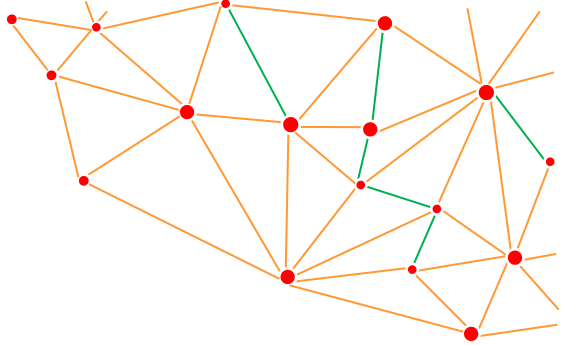
О, зная, зная, ти плачешъ, майко,
За туй, че ти си черпа робиня;
За туй, че твойтъ свѣщенъ гласъ майко,
Е гласъ-безъ помощъ, гласъ вовъ пустиня!

Плачи! Тамъ близо до градъ София
Вида азъ стърчи черно бѣсило.
И твойтъ *единъ* синъ Българю,
Виси на него. . . . Съсъ страшна сила



О, зная, зная, ти плачеш, майко,
затуй, че ти си черна робиня;
затуй, че твойт свещен глас, майко,
е глас без помощ, глас във пустиня.

Плачи! Там близо до град София
видя аз стърчи черно бесило.
И твойт един син, Българийо,
виси на него... със страшна сила.

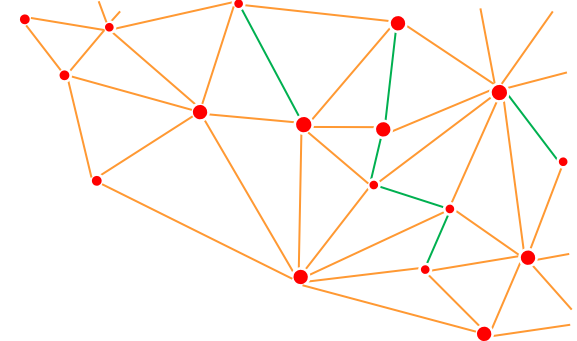


Откриване на грешки в текст

- Разпознаване на правописни, пунктуационни и граматични грешки в текст
- BERT (енкодер) модел
- Точността е 98 %

Документален филм разказва за извесния капитан Георги Георгиев. Той е първия българин извършил съмотно околосветско плаване с яхта. Преди това, участието му в едно от най-екстремните състезания по ветрохотство във света е одбелязано като рекорд. За мнозина капитана се превръща в символ на необикновена воля и духовна твърдос. Със своите успехи той прославя България.

Изводи



- КЛаДА-БГ е инфраструктура, която вече е започнала и ще продължи да работи не само с текст, но и с мултимодални данни - реч, изображения, видео, 3Д модели и др. -> **Знанието в КЛаДА-БГ се моделира в неговата многоизмерност, контекст и интегрираност.**
- В ерата на ИИ нарасна нуждата от (структурирани) качествени данни. -> **КЛаДА-БГ предоставя такива данни.**
- Макар че автоматичната обработка на данни вече се подпомага изключително много от ИИ, има нужда от добавяне на специализирано знание и знание за света -> **КЛаДА-БГ работи в тази посока чрез създаването на ИИ модели и RAG системи.**
- КЛаДА-БГ работи в европейски контекст, включително чрез членството си в два центъра за знание: *Център за южнославянските езици (Словения)* и *Център за големи езикови модели (Полша)*.