УНИВЕРЗИТЕТ У БЕОГРАДУ

Михаило Ђ. Шкорић

Композитне псеудограматике засноване на паралелним језичким моделима српског језика

докторска дисертација

Београд, 2022

UNIVERSITY OF BELGRADE

Mihailo Đ. Škorić

Composite pseudogrammars based on parallel language models of Serbian

Doctoral Dissertation

Belgrade, 2022

Ментори:

Проф. др Ранка Станковић, ванредни професор,
Универзитет у Београду, Рударско-геолошки факултет

доц. др Јелена Граовац, доцент,
Универзитет у Београду, Математички факултет

Чланови комисије:

Проф. др Владан Девецић, редовни професор,
Универзитета у Београду, Факултет организационих наука

доц. др Милош Утвић, доцент,
Универзитет у Београду, Филолошки факултет

Проф. др Драган Станков, ванредни професор,
Универзитет у Београду, Рударско-геолошки факултет

Датум одбране:_____

Наслов рада : Композитне псеудограматике засноване на паралелним језичким моделима српског језика

Алтернативни назив: /

Резиме: Циљ овог рада је да предочи предности коришћења композитних интелигентних система заснованих на паралелним архитектурама, а пре свега предност композитних псеудограматика заснованих на паралелним језичким моделима у обради, генерисању и евалуацији природног језика, и то поготово српског. У њему је најпре дат кратак увод у теорију формалних језика, предочене су различите врсте граматика и дат је преглед радова из области креирања њихових апроксимација. Описани су појмови псеудограматика и језичких модела и приказан је њихов историјски развој, са највећим акцентом на тренутно стање и најактуалније методе моделовања језика и језичке моделе. Уведена је проблематика евалуације квалитета текста, и описане су различите методе полу-аутоматске и аутоматске евалуације. У другом делу рада описана су два експеримента која су имала за циљ да утврде методологију креирања композитних система за потребе моделовања српског језика, при чему су описани начини креирања различитих репрезентација докумената и различити начини комбиновања излаза самосталних система у обради природног језика. Паралелни системи су том приликом успешно тестирани на задацима обележавања врста речи и утврђивања ауторства кроз моделовања мини-језика, где су остварили значајно боље резултате од самосталних метода. Коначно, описан је процес обучавања серије генеративних предобучених трансформера над различитим репрезентацијама корпуса српског језика и креирања композитних псеудограматика заснованих на тим моделима и различитим методама комбиновања. Развијени системи су евалуирани на задацима оцењивања квалитета текста, те проналажења и исправљања грешака. Приказани резултати издвојили су наслагани обучени класификатор као оптимални метод комбиновања језичких модела у јединствену псеудограматику.

Кључне речи: моделирање језика, језички модели, композитне структуре, машинско учење, српски језик, анализа текста, генерисање текста, аутоматска евалуација.

Научна област: Интелигентни системи

Ужа научна област: Обрада природног језика, Рачунарска лингвистика

Paper title: Composite pseudogrammars based on parallel language models of Serbian

Alternative title:/

Abstract: The aim of this paper is to present the advantages of using composite intelligent systems based on parallel architectures and, above all, the advantage of composite pseudogrammars based on parallel language models in the processing, generation, and evaluation of natural languages, especially Serbian. First a brief introduction to the theory of formal languages is given, distinct types of grammars are described an overview of papers in the field of creating their approximations were presented. The concepts of pseudogrammars and language models were described together with their historical development, with the emphasis on the current state-of-the-art and the best methods of language modelling and currently top-performing language models. The issue of quality evaluation of a text is introduced, and various methods of semi-automatic and automatic evaluation are described. In the second part of the paper, two experiments were described that aimed to determine the methodology of creating composite systems for the needs of modelling the Serbian language, where the ways of creating different representations of documents and diverse ways of combining the outputs of independent natural language processing systems were described. On that occasion, parallel systems were successfully tested on the tasks of part-of-speech tagging and authorship attribution through mini-language modelling, for which they achieved significantly better results than independent methods. Finally, the process of training a series of generative pretrained transformers on different representations of the corpus of the Serbian language and creating composite pseudogrammars based on those models and different combining methods is described. The developed systems were evaluated on the tasks of text quality evaluation and finding and correcting errors in the text. The presented results singled out the stacked trained classifier as the optimal method of combining language models into a unique pseudogrammar.

Keywords: language modeling, language models, composite structures, machine learning, Serbian language, text analysis, text generation, automatic evaluation.

Research area: Intelligent systems

Research subarea: Natural language processing, Computational linguistics

Садржај

I Увод

1
О језику и граматици

Како бисмо успешно могли да изучавамо природни језик и текст њиме писан морамо најпре тај исти језик подробније дефинисати и установити његове особине. Већ при том кораку јављају се проблеми јер, као и сваки други систем који је настао природно, еволутивно и без предумишљаја, и природни језик пати од исте бољке недостатка документације начина и сврхе, како настанка, тако и сопствене еволуције. Док код вештачких система изучавање може да отпочне уз процес изградње, код природних смо унапред закаснили, те смо осуђени на праћење трагова уназад и уклапање комплексних система у нове, нама познате и блиске оквире. Лингвисти који су се бавили темом аквизиције (усвајања) природног језика педесетих година двадесетог века су, са тим на уму, развили *теорију формалних језика* (Chomsky, 1956), засновану на ранијој идеји да се језик може описати математичким и логичким методама. Формални оквир који је тада зачет нашао је примену најпре у описивању структура, те препознавању и генерисању како постојећих, природних, тако и нових, вештачких језика.

1.1 Формални језици и граматике

Теорија формалних језика нам говори да се помоћу *формалне граматике* **G** може генерисати, као и препознати *формални језик* **L**(**G**), скуп исправних ниски (реченица) над алфабетом симбола неког језика**.** Формалну граматику у том смислу представља уређена четворка (**Σ, N, S, P**)**:**

- коначан скуп завршних симбола, *терминала*, који се не могу даље изводити – алфабет **Σ**;
- коначан скуп незавршних симбола, *нетерминала*, из којих се даље изводе други симболи – алфабет **N**;
- издвојени почетни (понекад зван и реченични) незавршни симбол **S**;
- коначан скуп изводних тј. трансформационих правила **P**.

Претпоставка теорије формалних језика је да се све реченице неког језика и само те реченице могу генерисати помоћу исправно конструисане *трансформационо-генеративне граматике,* која кроз поновљене примене изводних правила тј. *трансформација* елемената генерише нове, све сложеније и сложеније структуре (Harrison, 1978).

У претходно поменутом раду (Chomsky, 1956) дат је пример следеће, једноставне, трансформационо-генеративне формалне граматике **G** која генерише један опитни подскуп енглеског језика **L**(G):

Σ = {the man, took, the book}

N = {Noun Phrase, Verb Phrase, Verb}

S = {Sentence}

$\mathbf{P_1}$ = Sentence	⟶	Noun Phrase + Verb Phrase
$\mathbf{P_2}$ = Verb Phrase	⟶	Verb + Noun Phrase
$\mathbf{P_3}$ = Noun Phrase	⟶	the man, the book
$\mathbf{P_4}$ = Verb	⟶	took

Пример 1: формална граматика G која генерише један подскуп енглеског језика L(G). У прва три реда дефинисани су алфабети терминала и нетерминала, као и почетни симбол, а у наставку су дефинисана изводна правила, у којима се са леве стране налази неки незавршни симбол, а са десне стране симболи у које се он може трансформисати (раздвојени зарезом), а + представља размак између симбола, уколико их се, при трансформацији, добија више од један.

Граматика из примера развија се на следећи начин, како би се генерисале могуће реченице овог опитног језика:

1. Како је *Sentence* почетни симбол и једино правило за његово извођење је $\mathbf{P_1}$, оно ће бити коришћено приликом прве трансформације и тако ће се добити структура симбола *Noun Phrase* + *Verb Phrase***.**
2. Како ниједан од добијених симбола није терминал, мора се извршити даља трансформација. *Verb Phrase* се може даље трансформисати у још једну групу нетерминала путем правила $\mathbf{P_2}$ и тако ће се добити структура *Noun Phrase* + *Verb* + *Noun Phrase*.
3. *Noun Phrase* се даље може извести у два терминала *the man* и *the book*, док се *Verb* може извести једино у терминал *took*.

С обзиром на то да су након извршења четири трансформациона правила, сви елементи структуре терминали, трансформација се завршава, и можемо закључити да се језик **L**(G) састоји од укупно четири могуће реченице (два могућа *Noun Phrase* терминала**,** један могућ *Verb* терминал и два могућа *Noun Phrase* терминала → *2*1*2 = 4*):

1. the man took the man (*човек је узео човека*)
2. the man took the book (*човек је узео књигу*)
3. the book took the man (*књига је узела човека*)
4. the book took the book (*књига је узела књигу*)

Аутор напомиње да би се додавањем нових правила у скуп **P** скуп реченица могао проширити, а уколико би се укључила и рекурзивна правила, на пример, да реченице могу да садрже реченице (што наводи као чест случај са природним језицима), скуп би могао постати и бесконачан. Применом једног таквог рекурзивног правила, реченица *the man took the book*, би се могла наћи унутар друге, веће, реченице:

He thinks the man took the book.

Међутим, правила би се могла прилагодити и како би се, на пример, елиминисале реченице које нису у духу језика (Sinclair, 1984) попут *књига је узела човека*. Један од начина на који се ово може постићи јесте малим изменама у претходно описаној формалној граматици (Пример 1)—коришћењем служби (субјекат, предикат и објекат) уместо врста (именичке и глаголске синтагме) елемената у реченици. Нове елементе је потребно додати у алфабет нетерминала **N**, а можемо узгред додати и издвојену типографску капитализацију и интерпункцију у алфабет терминала **Σ**. Такође је неопходно, сходно томе, прилагодити и трансформациона правила.

Σ = {**The man**, took, the man, the book, **.**}

N = {**субјекат**, **предикат**, **објекат**}

S = {Sentence}

$\mathbf{P_1}$ = Sentence → **субјекат + предикат + објекат + .**

$\mathbf{P_2}$ = **субјекат** → The man

$\mathbf{P_3}$ = **предикат** → took

$\mathbf{P_4}$ = **објекат** → the man, the book

Пример 2: формална граматика G_2 која генерише подскуп енглеског језика $L(G_2)$, унапређену верзију претходно описаног језика (Пример 1).

Ова граматика, $\mathbf{G_2}$, развија се на следећи начин:

1. Како је и у овом примеру *Sentence* почетни симбол, а једино правило за његово извођење (поново) $\mathbf{P_1}$, оно ће бити коришћено приликом прве трансформације и тако ће се добити структура *субјекат* + *објекат* + *предикат* + „*тачка*“**.**
2. Пошто након овог корака и даље имамо три нетерминала користићемо остала правила како би се граматика даље развила. Примењивањем правила $\mathbf{P_2}$, $\mathbf{P_3}$ и $\mathbf{P_4}$ добијамо структуру која се састоји искључиво од терминала и тиме се развијање завршава.

На основу овога можемо коначно закључити да се нови језик $\mathbf{L(G_2)}$ састоји од укупно две могуће реченице (један могући субјекат, један могући предикат, два могућа објекта и тачка на крају реченице → *1*1*2*1=2*):

1. The man took the man. (*Човек је узео човека.*)
2. The man took the book. (*Човек је узео књигу.*)

1.2 Усвајање језика код човека и универзална граматика

Према теоретичарима формалних језика човек поседује јединствену структуру коју користи како би употребом познатог лексикона и правила која над њим важе, *његове личне трансформационо-генеративне граматике*, генерисао (а и препознавао претходно генерисане) реченице у циљу комуникације. Дакле, човек који се служи неким природним језиком може да склапа реченице, као и да одреди да ли нека реченица припада том језику или не (Wexler & Culicover, 1980).

У раду који је настао кооперацијом биолога, психолога и лингвиста (Hauser, et al., 2002) испитује се хипотеза да је човек еволутивним путем дошао до својеврсног *органа за усвајање језика* (у виду структуре неурона у мозгу), који му је касније омогућио усвајање сложеног система комуникације којим се и данас користи. Иако многе животињске врсте имају способност комуникације између јединки, оно што је у раду потенцирано као јединствено својство овог система код човека је могућност употребе рекурзије за склапање и разумевање сложених реченица, и то без употребе сувишне количине радне меморије у којој би њихово значење било забележено. Хипотеза је да је висока награда за коришћење тог система потом довела до тога да се еволутивно распространи на читав људски род. Даље, према овој теорији, сваки човек поседује сопствену граматику коју је усвојио путем комуникације са другим људима, и која је исправно генерисана над његовим органом за усвајање језика. Са друге стране, хипотетичка граматика чија својства дели одређена група људи генерише природни језик којим ти људи говоре, на пример, *српски језик*.

По узору на то, у класичној лингвистици граматика природног, људског језика се базира на ограниченом броју опажених реченица (корпусу) и формулише се на основу оквира дефинисаних путем претходно поменуте теорије формалних језика, где се елементи формалне граматике која је предмет дефинисања изводе коришћењем метода *корпусне лингвистике.* Из корпуса се најпре изводи скуп

терминала—алфабет **Σ** (који представља *лексикон* речи опаженог језика), потом скуп нетерминала—алфабет **N,** који представља скуп свих граматичких категорија којима речи из алфабета **Σ** припадају, те коначни скуп изводних правила **P,** према којима су конструисане све прикупљене реченице (Илустрација 1). Прикупљени скупови се затим пројектују на нови, неограничени домен реченица који, по теорији, *треба* да представља тај језик, и само тај језик (Sinclair, 1991). Под предусловом да за језик постоје јасни скупови граматичких и неграматичких реченица успех израђене граматике се може и експериментално тестирати.

Илустрација 1: Процес креирања граматике природног језика коришћењем метода корпусне лингвистике и корпуса текстова писаних на том језику.

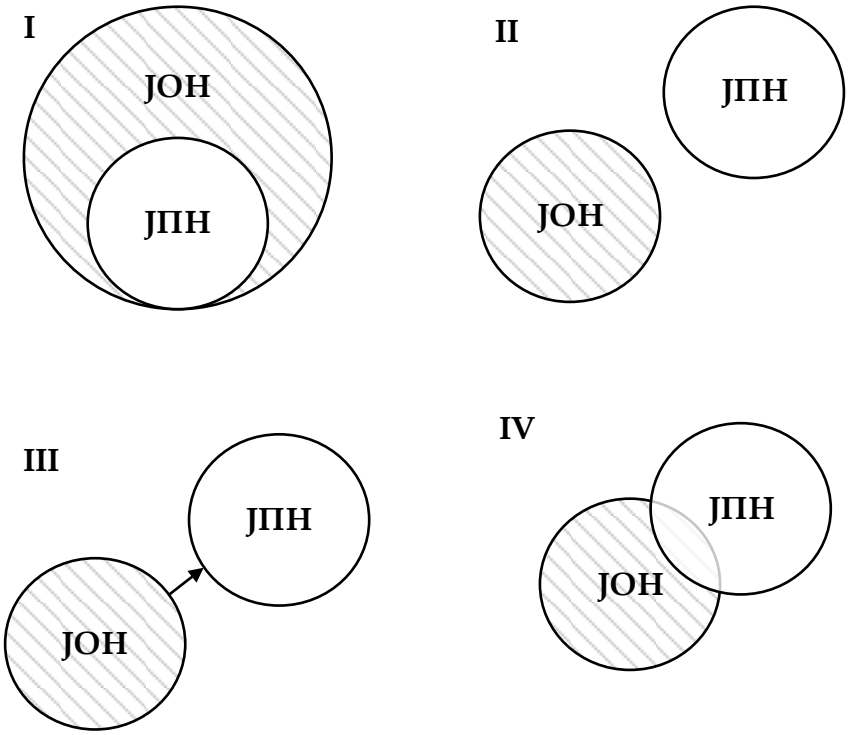
Један од проблема при креирању корпуса одређеног језика је свакако проблем граматичке хетерогености његових корисника и генералне несагласности око граница између сродних језика уколико оне уопште постоје и уколико су јасне (Chomsky, 2013). Чомски примећује да су се, пре него што су школе преузеле подучавање становништва језиком и пре него што су установљене националне граматике, језици глатко преливали и није постојала јасна граница између сродних језика, на пример, каталонског и француског. Из тога произилази закључак да, пре свега, не постоји коначан број формалних језика, те да се они међусобно преклапају, мешају и преливају, као и да је вероватнија тврдња да постоји једна, универзална граматика (Lyons, 1981), крајњи продукт органа за усвајање језика, са способношћу да генерише све могуће реченице свих језика. То ипак не искључује могућност да постоји рекурзивно конструисан, бесконачан број мањих граматика које генеришу различите, како признате, природне језике којима се људи служе, тако и вештачки конструисане језике који су повезани универзалном граматиком на вишем нивоу.

1.3 Језици и граматике посебне намене

Подела језика игра важну улогу у његовом дефинисању, али и схватању. Поред поделе на природне и вештачке, важна је и она према њиховој употреби. Језици за општу употребу, такозвани *језици опште намене (language for general purpose),* у даљем тексту ЈОН, попут *српског,* или *пољског* језика, немају специфичну употребу, осим те да су коришћени за свеопште споразумевање са другим особама које га користе. Са

друге стране, *језици посебне намене (language for specific purpose)*, у даљем тексту ЈПН, везани су за употребу у одређеном домену или друштвеној сфери, а примери оваквих језика су *научни српски* језик или *правни пољски* језик.

Иако су у теорији широко примењиви и примењени, дефинисање ЈПН је само по себи проблематично. Није јасно шта су они тачно, које су им границе, као ни њихова веза са ЈОН. Намеће се тако више приступа у њиховом дефинисању (De Beaugrande, 1987)(Илустрација 2):



Илустрација 2: Визуелизација четири приступа дефинисања односа између језика опште намене – ЈОН и језика посебне намене – ЈПН.

I. Први приступ у покушају дефинисања ЈПН био је да се он означи као стилски подскуп ЈОН, из којег је хипотетички и настао. Овај приступ узима за основну претпоставку то да ЈОН садржи у себи ЈПН. Ипак, адекватни докази за ово нису пронађени у тренутку настајања ове теорије. Штавише, експерименти су показали да се стил говора или језик говорника различитих ЈПН често подједнако разликује као и језик говорника различитих ЈОН.
II. Други приступ је уследио као одговор на I приступ и имао за идеју да се сваки од ЈПН-а дефинише као засебан језик са засебним лексиконом и граматичким правилима. По томе се ЈОН и ЈПН не преклапају, те би *научни српски* језик био обележен као засебни природни језик независан од *српског* језика. Иако се у овом приступу полемише да ЈПН често имају засебан лексикон без чијег познавања може постати неразумљив регуларним говорницима, као проблем се намеће то да он већину ресурса дели са сродним ЈОН и граматичка правила помоћу којих се генерише нису засебно дефинисана, као што је раније поменуто. Други проблем овог приступа је да су речи специфичне за ЈПН често

интернационалне и користе се у другим ЈПН из истог употребног домена, што додатно оспорава њихову независност.

III. Трећи приступ је био дефиниција ЈПН као вештачки настале компоненте ЈОН. Према овоме, ЈОН је настао природним путем, док је ЈПН настао из њега касније, вештачким путем. Овај приступ, као и сваки који користи сличну поделу (на природно и вештачко) пати од бољке непрецизне дефиниције природног језика.

IV. Четврти приступ дефинише ЈПН искључиво на основу његове употребе. ЈПН је дефинисан као унија свих језичких јединица које се користе у специфичне комуникационе сврхе. Овај приступ наочиглед решава проблеме претходна три, а његове методе се ослањају у потпуности на корпусну лингвистику, где се лексикон и правила ЈПН екстрахују из корпуса текстова специфичне употребе, као што је пракса и у савременој лингвистици. Кроз експерименте рађене овим приступом дошло се до закључка да се ЈОН и ЈПН делимично преклапају, радије него да један садржи други, што додатно подупире претходно поменуту хипотезу да не постоји јасна граница између природних језика (Chomsky, 2013), па тако ни између ЈОН и ЈПН. Из тога, међутим, произилази главни проблем са којим се сусреће овај приступ, ако не у дефиницији, онда у методологији: прецизна дефиниција специфичних употреба је немогућа, па тако и сакупљање, на пример, искључиво *научног* или искључиво *правног* корпуса неког језика.

1.4 Типови језика и граматика према начину генерисања

У теорији формалних језика, онако како ју је видео Чомски, граматике језика су према начину генерисања језика подељене у четири групе, смештене у угњеждену хијерархију (Chomsky, 1956) (где свака хијерархијски виша група садржи све хијерархијски ниже групе) коју називамо и *хијерархија Чомског* или *хијерархија Чомски-Шуценберга* (Илустрација 3).

Илустрација 3: Популаран приказ типова језика према хијерархији Чомског.

Сваки од ова четири типа граматика одговара одређеном типу језика (обележеним природним бројем од 0 до 3) који те граматике генеришу, са тим да треба имати у виду да хијерархијски више граматике (чији је тип обележен мањим бројем) могу да генеришу и све хијерархијски ниже језике. Граматике најнижег типа, типа 3, генеришу најпростије, регуларне језике. Контекстно-слободни језици захтевају граматике типа 2, контекстно осетљиви језици захтевају граматике типа 1, док се рекурзивно-пребројиви језици могу генерисати само помоћу граматика највишег реда, граматика типа 0.

Свакој од поменутих граматика одговара и одређени аутомат који, са становишта рачунарства, има могућност да препознаје исте језике као та граматика (Chomsky & Schützenberger, 1959). Најпознатији такав аутомат је апстрактна, хипотетичка *Тјурингова машина* која, у теорији, коришћењем бесконачног приступа меморији може да изврши било који валидан рачунарски алгоритам, а коју је 1936. године конципирао Алан Тјуринг (Turing, 1938). На тај начин, узевши у обзир да свака граматика G генерише један и само један језик L(G), можемо и тестирати да ли нека ниска задатих симбола припада том језику помоћу рачунарског аутомата који је еквивалентан граматици (Jager & Rogers, 2012):

Граматике типа 0, назване такође и граматике без ограничења, генеришу све формалне језике које може да препозна Тјурингова машина, а тако и обратно, било који алгоритам који се може извршити на Тјуринговој машини може се написати у облику граматике типа 0. Правила извођења нису ограничена и форме су $\boldsymbol{\alpha} \rightarrow \boldsymbol{\beta}$, где $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$ могу бити било ког облика. Примери језика које ове граматике генеришу су рачунарски програми који се не извршавају у недоглед, већ имају коначан број корака или услов након којег извршавање престаје. Уколико извршавање алгоритма на Тјуринговој машини траје бесконачно, то значи да задати унос (ниска симбола) не припада језику који генерише одговарајућа граматика.

Граматике типа 1, контекстно-осетљиве граматике, генеришу истоимени тип језика, а аутомат који им одговара је *линеарно-ограничени аутомат* (врста *недетерминистичке Тјурингове машине* са условима ограничења). Као и одговарајући аутомат, правила граматике садрже ограничења, наиме, свако правило извођења тежи томе да упрости структуру реченице C, те за свако правило важи да ниска на левој страни правила не сме бити дужа од ниске на десној страни правила ($\boldsymbol{\alpha A \beta} \rightarrow \boldsymbol{\alpha \gamma \beta}$, где је $\boldsymbol{A}$ нетерминал, а $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ и $\boldsymbol{\gamma}$ су подскупови уније алфабета терминала и нетерминала). Поменуто ограничење омогућава, за разлику од граматика типа 0 (Тјурингове машине), да се за произвољан унос може у коначном времену одредити да ли припада језику, при чему иако коначно, време препознавања уноса, услед комплексности проблема, може бити произвољне величине. Примери контекстно-осетљивих језика су скуп природних бројева (где је сваки број n представљен ниском дужине n) као и скуп квадрата природних бројева.

Граматике типа 2, контекстно-слободне граматике, генеришу контекстно-слободне језике, а правила извођења морају бити облика $\boldsymbol{A} \rightarrow \boldsymbol{\beta}$, где је $\boldsymbol{A}$ један нетерминал, а $\boldsymbol{\beta}$ ниска над унијом алфабета терминала и нетерминала. Аутомат који одговара овим

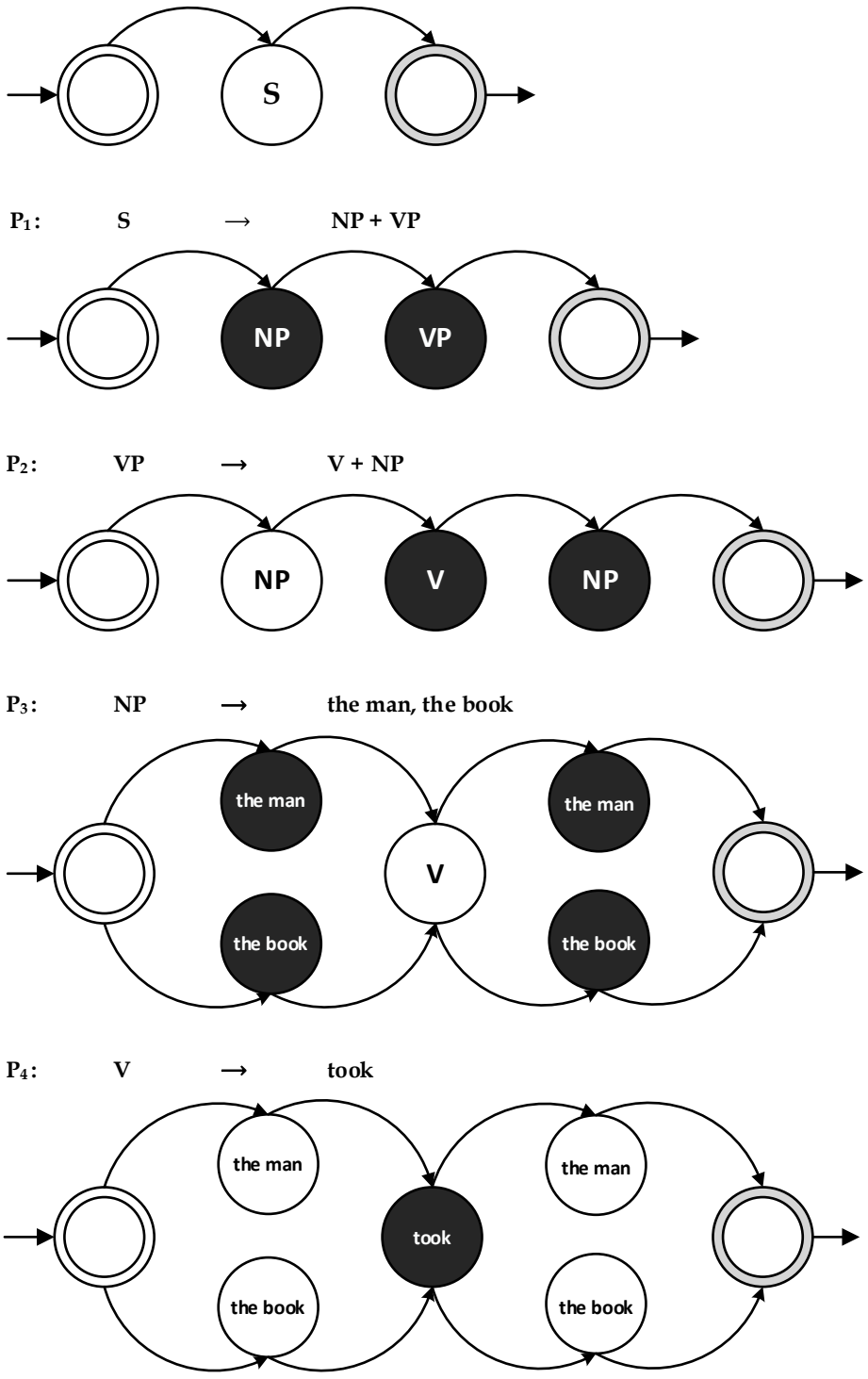
граматикама је *недетерминистички потисни аутомат*, а примери контекстно-слободних језика су програмски језици вишег реда, као и аритметички изрази.

Граматике хијерархијски најнижег типа—типа 3, зову се такође и регуларне граматике и генеришу регуларне језике. Граматике тип 3 могу бити десно-линеарне (правила извођења су облика или $A \rightarrow a$ или $A \rightarrow aB$, где A и B представљају нетерминале, а a терминал) или лево-линеарне (правила извођења су облика $A \rightarrow a$ или $A \rightarrow Ba$). Језици који се препознају овим граматикама одговарају онима који се препознају путем *коначних аутомата*, а који се могу описати и *регуларним изразима*. Регуларне граматике (а самим тим и коначни аутомати и регуларни изрази) се најчешће користе за дефинисање претрага текста, као и за описивање мањих подскупова природних језика. Сваки коначан језик је регуларан, као и језик добијен применом коначног броја скуповних операција (унија, пресек, разлика), операције дописивања и операције Клинијевог затворења над регуларним скуповима.

Коначни аутомат, који одговарају одређеној регуларној граматици G, представља математички модел машине која за задату улазну ниску одговара са *да* или *не* на питање да ли она припада језику генерисаном регуларном граматиком G. Коначни аутомат се дефинише преко коначног скупа могућих стања, међу којима се нека могу издвојити као почетна и завршна, као и преко таблице прелаза којом се описује прелазак аутомата из једног стања у друго у зависности од текућег стања аутомата и прочитаног симбола улазне ниске. Аутомат у сваком тренутку може бити само у једном стању.

Међу различитим типовима дефиниција коначних аутомата посебно место имају оне код којих увек постоји тачно једно почетно и једно завршно стање. У почетном стању аутомат почиње са радом и чита први симбол улазне ниске и на основу таблице прелаза мења текуће стање или остаје у истом стању, а онда се процес понавља за сваки следећи прочитани симбол улазне ниске. Ако аутомат прочита целу улазну ниску и после тога текуће стање буде завршно, аутомат је препознао улазну ниску; у сваком другом случају (аутомат није прочитао целу улазну ниску и не може да настави са радом или је по читању целе улазне ниске текуће стање незавршно) улазна ниска није препозната. Промена стања, од почетног преко прелазних стања све до завршног стања назива се *транзиција*, а она се одвија према задатом скупу правила која одговарају правилима регуларне граматике (Carroll & Long, 1989).

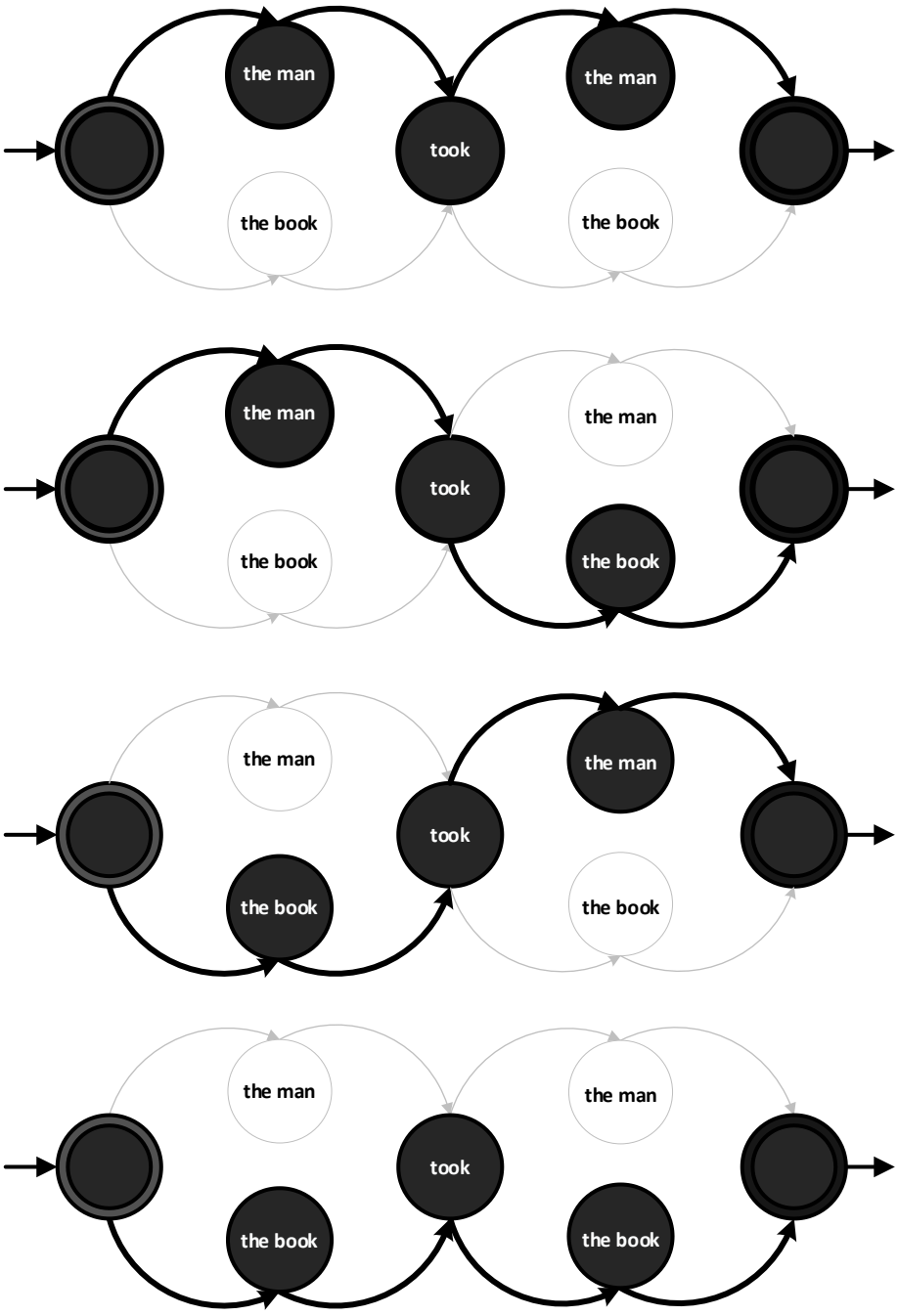
Тако се, на пример, граматика опитног подскупа енглеског језика (Пример 1), може описати и путем коначног аутомата (Илустрација 4), где је коначни аутомат који одговара тој граматици приказан на дну илустрације, и добија се применом последњег правила трансформационо генеративне граматике.



Илустрација 4: Развијање прелазних стања коначног аутомата, према елементима и правилима описаним кроз Пример 1, а које опонаша развијање формалне граматике G из истог примера, где *S* означава реченицу, *NP* означава компоненту *Noun Phrase*, *VP* означава компоненту *Verb Phrase*, а V означава компоненту *Verb*.

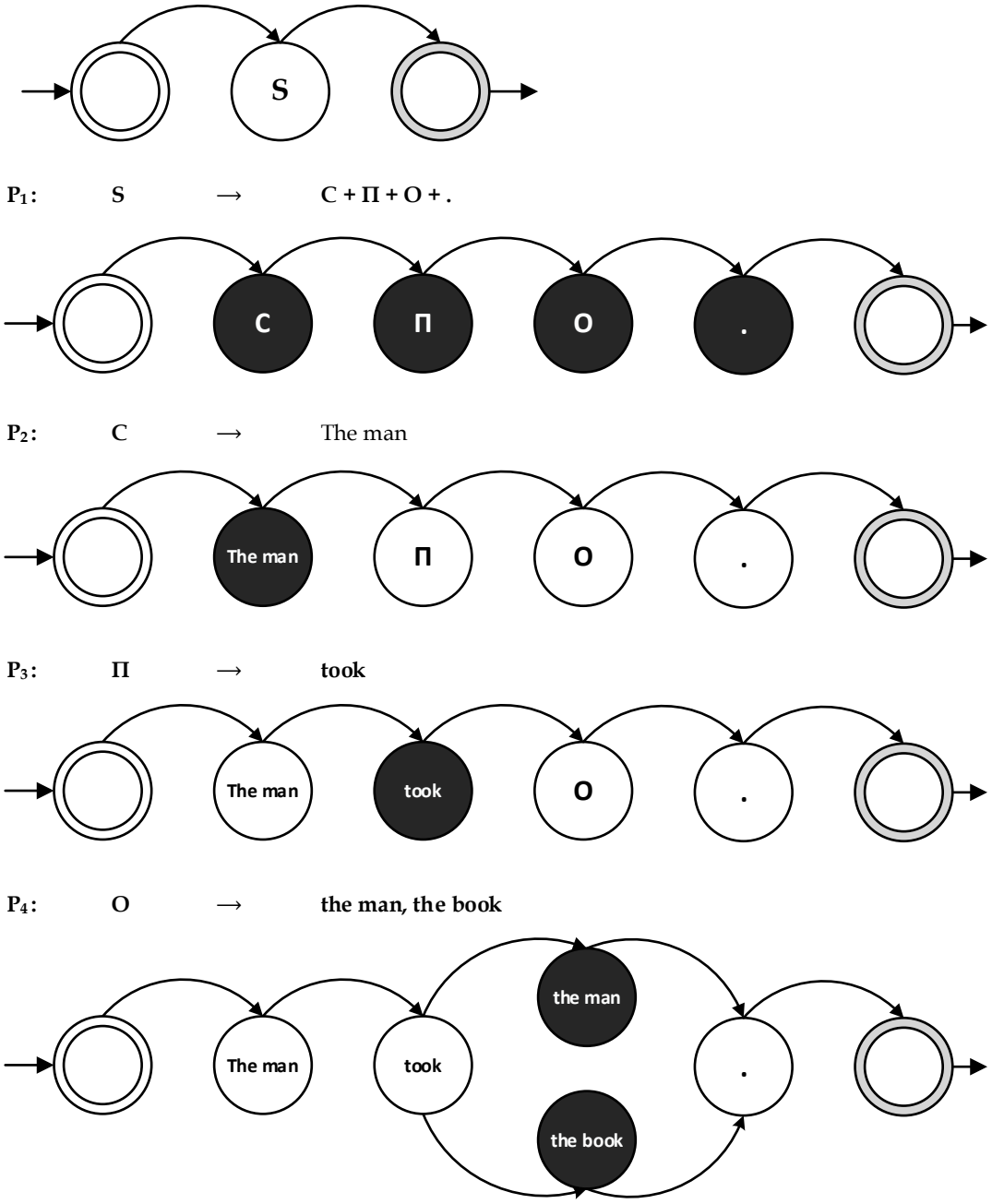
На датој илустрацији, а у сваком реду, први чвор са лева представља почетно стање и последњи представља коначно стање, док чворови између представљају прелазна

стања (при чему су у сваком реду нови чворови добијени трансформацијом обојени у црно). Конструкцијом овог аутомата смо уједно показали и да је граматика коју је Чомски дао као пример регуларна, и да генерише један регуларан језик. Генерисање свих могућих реченица над тим аутоматом се одвија кроз све могуће комбинације његових транзиција над прелазним стањима (Илустрација 5).

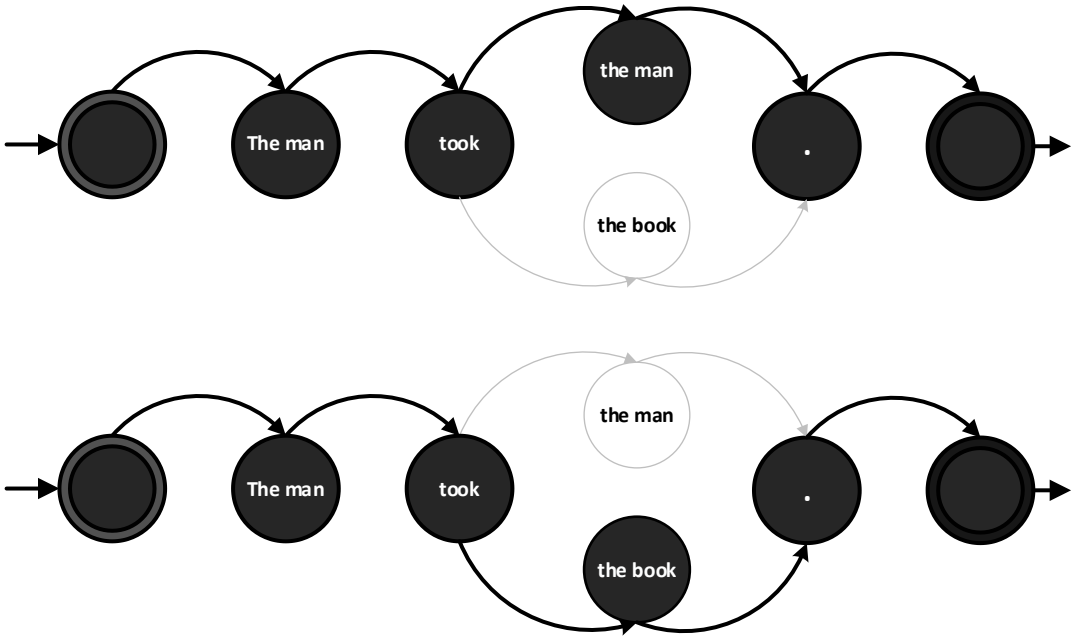


Илустрација 5: Могуће комбинације транзиција (обојене у црно) за претходно развијени коначни аутомат за граматику G, које генеришу све могуће ниске терминала над њим: *the man took the man*, *the man took the book*, *the book took the man* и *the book took the book*.

На исти начин на који је приказано развијање аутомата над правилима формалне граматике G (Илустрација 4), можемо приказати и развијање аутомата над правилима формалне граматике G_2 (Пример 2) (Илустрација 6), као и генерисање реченица над тим аутоматом (Илустрација 7).



Илустрација 6: Развијање прелазних стања коначног аутомата, према елементима и правилима описаним кроз Пример 1, а које опонаша развијање формалне граматике G_2 из истог примера, где *S* означава реченицу, *С* означава субјекат, *П* означава предикат, а О означава објекат.



Илустрација 7: Могуће комбинације транзиција (обојене у црно) за претходно развијени коначни аутомат за граматику G_2, које генеришу све могуће ниске терминала над њим, а то су само две реченице: *The man took the man.* и *The man took the book.*.

Упркос овим примерима, и даље је упитно у који тип језика (према хијерархији Чомског) спадају природни језици.

1.5 Регуларност природних језика

Иако је дуги низ година примарни метод описа природних језика био путем регуларних граматика, постоје ваљани аргументи зашто природни језици нису регуларни, а за пример можемо узети неко од истраживања која кажу да контекстно-слободне граматике боље генеришу природне језике (Marcus, 1997). До сличног закључка се долази и приликом израде формалне граматике српског језика (Đorđević, 2017), где се ослања на граматику адјунгованих стабала, која одговара контекстно-слободним граматикама у хијерархији Чомског. Са друге стране, постоје и неки аргументи зашто природни језици јесу регуларни (Kornai, 1985):

- С обзиром на то да је једини начин да се покаже да природни језици нису коначни јесте да се покаже неки бесконачан низ граматичких реченица, то просто није могуће доказати (ако се не узимају у обзир рекурзивне реченице).
- Будући да је људски мозак заснован на неуронима, и да се појединачни неурони могу моделовати коначним аутоматима (McCulloch & Pitts, 1943), а коначни тродимензионални низ таквих аутомата може бити замењен једним аутоматом (Kleene, 1956), природни језици морају бити регуларни, ако их можемо обрађивати у нашем мозгу.

Но чак и када би се испоставило да природни језици јесу регуларни, па чак и контекстно-слободни, граматике које би их прецизно описивале (а поготово генерисале) имале би тако велики број правила да се њихово записивање данас сматра непрактичним задатком (Karttunen, et al., 1996).

2
О псеудограматикама и језичким моделима

2.1 Псеудограматике и језички модели

У пракси је установљено да је један од највећих недостатака формалних граматика велика цена њиховог настанка. Извлачење граматичких правила из корпуса текстова се може, наравно, вршити њиховим пописом, али при томе долази до проблема превелике конкретизације модела (*overfitting*), где се појединачна правила узимају за општа и губи се шира слика. Са друге стране, деривација општих правила из појединачних мора се обављати пажљиво и изискује огромну количину времена.

Крајем двадесетог века, са убрзаним развојем технологија вештачке интелигенције (а посебно метода машинског учења), распламсала се идеја да се на бржи начин може доћи до доброг резултата у многим сферама, где је ручно дефинисање правила изискивало превише времена. То је утицало и на корпусну лингвистику, где су истраживачи почели да примењују нове технологије, понајпре на аутоматизацију екстракција граматика из текстуалних корпуса. Тај задатак се убрзо разлио и у израду потпуно нових аутомата заснованих на вероватноћама, који емулирају аутомате и граматике засноване на правилима. Те нове системе, који уместо буловског одговора нискама додељују вероватноће на основу претходно опаженог корпуса текстова, називамо *језичким моделима* (Илустрација 8), док језичке моделе који уз способност препознавања имају и способност генерисања језика називамо *псеудограматикама* (Mizumoto, et al., 1972; Geman & Johnson, 2002).

У случају језичких модела, ниске се најчешће односе на реченице или ниске *токена*, градивних делова реченица који се могу наћи у форми н-грама речи или карактера (који одговарају терминалима у теорији формалних језика). Дакле, језичке моделе

дефинишемо као системе који додељују вероватноће нискама на основу контекста у којем се јављају, а те системе моделирамо на основу претходно прикупљеног корпуса.

Илустрација 8: Грубо поређење функционалности формалне граматике (горња трака) и језичког модела (доња трака) неког језика *L*, где је *S* нека ниска токена, а *P(S∈L)* вероватноћа да ниска *S* припада језику *L*.

Претходних неколико деценија моделовање језика се развијало пре свега у два правца: моделовање засновано на н-грамским статистикама и моделовање засновано на вештачим неуронским мрежама, према инспиративној идеји Елмана (Elman, 1988; Elman, 1990) који је, док се бавио временским серијама као улазним подацима за моделе машинског учења, конструисао вештачку неуронску мрежу чији је циљ био да предвиди наредни елемент у низу.

2.2 *Моделовање језика засновано на вештачким неуронским мрежама*

Најзначајнији напредак у моделовању језика коришћењем вештачких неуронских мрежа (у даљем тексту ВНМ) забележен је крајем двадесетог века у часопису *Neural computation* Универзитета у Кембрицу, где је у свега неколико година (од 1989. до 1992.) објављена серија радова на тему апроксимација коначних аутомата и формалних граматика коришћењем ВНМ.

Први у серији радова важних за апроксимацију граматика објавили су истраживачи са Карнеги-Мелон Универзитета (Cleeremans, et al., 1989). Они су изучавали способност ВНМ да опонаша коначни аутомат—регуларну граматику G језика L(G), уколико је обучена на одговарајућем узорку реченица тог језика. Након експеримената на неколико различитих припремљених корпуса, закључено је да оваква мрежа одговара коначном аутомату који производи бесконачан број реченица регуларног језика, након што је обучена на коначном скупу реченица из корпуса за обучавање.

Истраживачи са Универзитета у Остину објављују свој напредак на пољу апроксимације две године касније (Park & Sandberg, 1991). За разлику од колега, они се фокусирају на проблем универзалне апроксимације, путем које се уједно може апроксимирати и граматика природног језика. Њихов приступ, заснован на

радијалној (*radial basis function*) активацији као алтернативи за тада стандардну сигмоидну активацију неурона, производи свеукупно боље резултате у универзалној апроксимацији.

У кооперативном раду који су објавили истраживачи са Универзитета у Принстону и Универзитета у Мериленду (Giles, et al., 1992), показано је да посебна врста ВНМ— рекурентна неуронска мрежа вишег реда—са статистичком значајношћу опонаша аутомат мањих регуларних граматика, са дужином ниски мањом од сто карактера. Такође, они су успешно направили пресликавање које се може користити за трансформацију неуронске мреже у коначни аутомат и обратно, са стопостотном прецизношћу.

Резултати сличног експеримента са рекурентним неуронским мрежама другог реда објављени су у истом издању часописа (Watrous & Kuhn, 1992), са разликом да су за обучавање коришћени позитивни и негативни примери, а циљ је био да се аутоматски направи коначни аутомат који са сигурношћу потврђује или негира да улазна ниска припада унапред одређеном језику. Експеримент је био успешан на свим примерима, са тим да је примећено да фина подешавања функција унутар самог система имају значајан утицај на добијене резултате.

Јасна веза између формалне граматике и њене апроксимације, псеудограматике, потврђује се и касније у још једном, другачије постављеном истраживању (Tsoulos, et al., 2008). У експерименту који је спроведен, аутори користе *граматичку еволуцију*, генетски алгоритам заснован на машинском учењу, са циљем да произведу *еволутивну неуронску мрежу*. ВНМ креирана за потребе овог експеримента (укључујући и топологију и вредности параметара) записана је у облику формалне граматике и развија се користећи њена трансформациона правила, а аутори су показали да је пресликавање у оба смера могуће и код овог примера.

2.3 *Н-грамски-засновано моделовање језика*

Као што је претходно приказано (Илустрација 8), задатак језичких модела је додељивање вероватноћа нискама на основу неког контекста. Ипак, истраживачи су у пракси приметили да је број различитих могућих контекста превелик да би се могао успешно моделовати, тј. да би се вероватноће могле израчунати и доделити за било коју ниску. Дошло се до тога да би смањивањем контекста у затворене н-грамске структуре, омогућило догледно израчунавање, па би се, уместо свих претходних токена контекст ограничио, на пример, на само последњи (биграмски модел) или последња два (триграмски модел), тако би се, на пример, вероватноћа неке речи израчунавала у односу на две претходне. Треба напоменути и да се осим за ограничавање контекста н-грами могу користити и за груписање низа токена којем желимо да доделимо вероватноћу, но ова операција је скупа, јер се време израчунавања готово експоненцијално повећава па су такви модели ретки (Manning & Schutze, 1999).

Што се тиче поједностављивања израчунавања путем смањивања домена контекста, још један популаран начин је *класно-засновано моделовање* језика (Brown, et al., 1992). При коришћењу овог приступа токени сличних значења или блиских граматичких обележја се групишу у класе, а онда се те класе (уместо самих токена) користе приликом израчунавања вероватноћа, што би било еквивалентно обрађивању над нетерминалима у формалним граматикама. Осим путем правила, за шта је неопходна не само људска, већ и експертска интервенција, погодне класе се могу дефинисати и путем метода машинског учења, што је веома захтеван процес са становишта рачунарства. Дакле, класно-засновано моделовање омогућује догледно израчунавање вероватноћа, али захтева скупљу припрему језичког модела.

Приликом израчунавања вероватноћа могу се користити додатне мере за унапређење тачности модела, на пример, засноване на тежинској измени вероватноћа. Један од начина је такозвано *кеширање*, метода где се, приликом израчунавања, додељује већа вероватноћа нискама које садрже н-граме који се већ појављују у контексту. Овај приступ је заснован на претпоставци, да ако се нека ниска већ једном појавила у тексту, постоје повећане шансе да ће се она појавити поново (Kuhn & De Mori, 1990).

Друга популарна метода је прилагођавање модела одређеном задатку, такозвана *дискриминација модела*. Тежине за одређене н-граме се проналазе углавном путем *дообучавања* (*fine tuning*) (Chen, et al., 2000) које се обично врши на засебном скупу посебно припремљених реченица или н-грама токена. Алтернативно, приликом креирања основног корпуса који ће послужити за моделирање, те реченице могу бити додате тада, али вишеструко, како би додатно утицале на процес креирања модела.

2.4 *Савремено моделовање језика*

Премда је потенцијал коришћења ВНМ за моделовање језика уочен рано, ограничења која овај приступ намеће су учинила да се развој привремено успори. Велика количина података за обучавање неопходна за правилну генерализацију граматичких правила, као и задовољавајући рачунарски ресурси (поготову у виду количине радне меморије и процесорске моћи) нису били доступни (бар не широј јавности) у тренутку развијања методологије, што је узроковало поменути пораст у коришћењу једноставнијих, н-грамских метода (које су се уз то показале и прилично успешне за обављање одређених задатака). Додатно, у пракси је примећен и проблем *нестајућег градијента,* који је последица пропагирања градијента грешке уназад (*backpropagation*) при обучавању вишеслојних и рекурентних ВНМ (Hochreiter, 1991), а овај проблем се посебно истицао при моделирању природног језика.

2.4.1 Дубоко учење

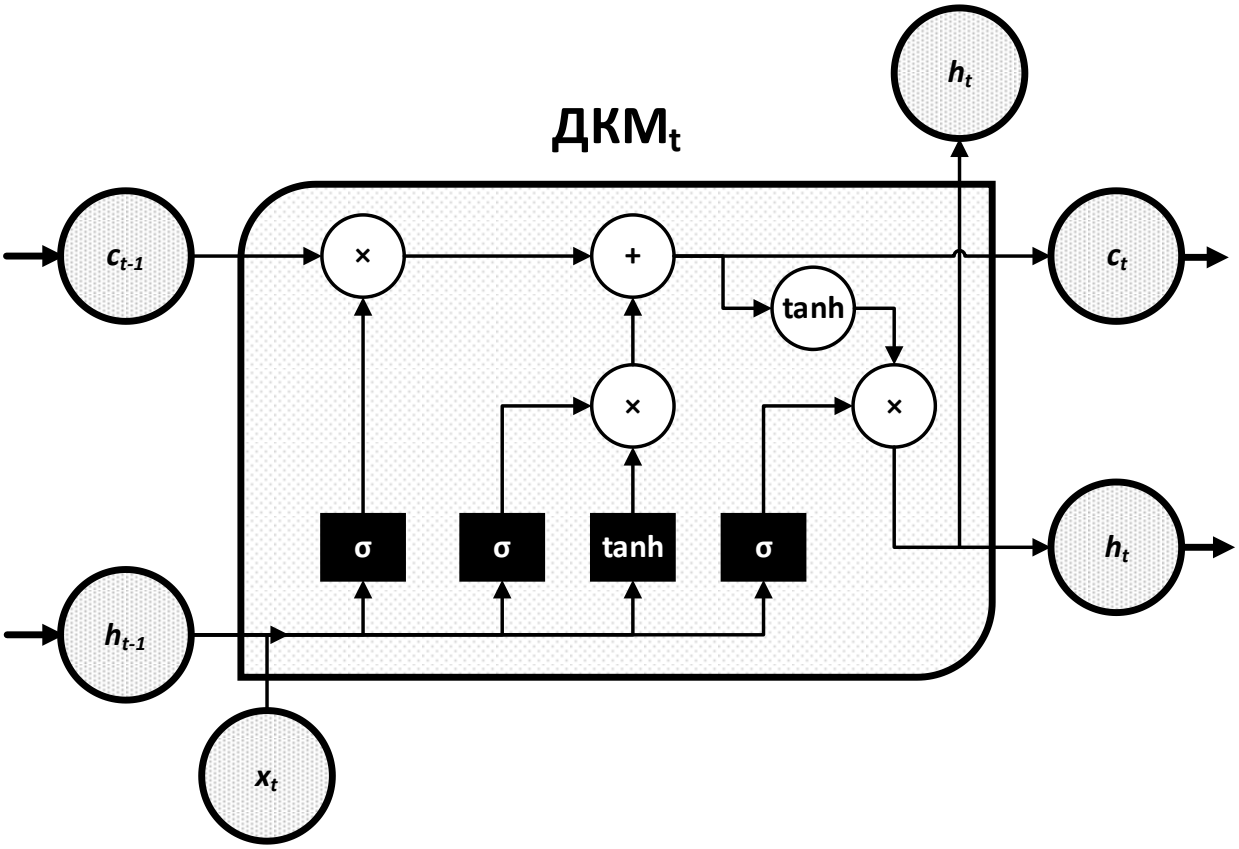
Ипак, експоненцијални раст рачунарске моћи доступних рачунара који је потом уследио, као и експоненцијални пораст количине доступних (између осталог, текстуалних) података у оквиру феномена *Big Data*, омогућио је да теорија коначно буде и технолошки подупрта, покренувши нови талас свежих истраживања,

заснованих на идеји *дубоког учења* (*deep learning*) (LeCun, et al., 2015), тренутно најзаступљенијој подобласти истраживања машинског учења, и вештачке интелигенције уопште (Илустрација 9).

Илустрација 9: Хијерархијски однос поља вештачке интелигенције, машинског учења и дубоког учења.

Употреба метода *long short-term memory* (дуге краткорочне меморије, у наставку текста ДКМ) (Hochreiter & Schmidhuber, 1997), на први поглед је решавала проблем нестајућег градијента, а уз то и пружала боље, до тада недостижне резултате. Изабрани неурони ВНМ замењени су ДКМ скривеном ћелијом (Илустрација 10), која процесира и селективно пропушта улазе, омогућујући пропагирање информација дубље у мрежу. Ипак, ово процесирање се испоставља као веома скупо, јер обучавање ВНМ са ДКМ ћелијама траје много дуже, а уз то и не решава у потпуности проблем нестајућег градијента, који постаје видљив стагнирањем резултата када се додају додатни и додатни слојеви на постојећу ВНМ са ДКМ скривеним ћелијама.

Исти метод, када је примењен у моделовању природног језика (Sundermeyer, et al., 2012), убрзо постаје стандард у високо-профилном моделовању језика (поготову за потребе озбиљних пројеката) упркос представљеним ограничењима, а наравно, само ако су рачунарски ресурси доступни. Тек са појавом *Трансформерске* архитектуре (Vaswani, et al., 2017), коју је развио *Гугл* (*Google*), као адекватне алтернативе за моделе са ДКМ меморијском ћелијом, направљен је нови искорак на пољу моделовања природног језика.



Илустрација 10: Структура ДКМ меморијске ћелије, где је x улазни вектор, *h* излазни вектор, *c* вектор стања (*t* је тренутна ћелија, а *t-1* претходна), σ представља јединствен слој перцептрона, а *tanh* хиперболични тангенс. Стање ћелије, као и њен излазни вектор се користе као додатни улази за следећу ДКМ ћелију у низу.

Основна разлика између трансформера и ВНМ са ДКМ је то што се трансформери не ослањају на рекурентне структуре, већ имају побољшан модел *пажње* (*attention*), специјалног параметра који се пропагира током учења, а који служи одвајању битних од небитних информација. Данас су најзначајнији и најзаступљенији језички модели засновани управо на овој архитектури, која се базира на *енкодер-декодер* структури за обучавање, подупртој предобученим (*pre-trained*) *векторизацијама* речи (*word embeddings*), и која се реализује на следећи начин (Илустрација 11):

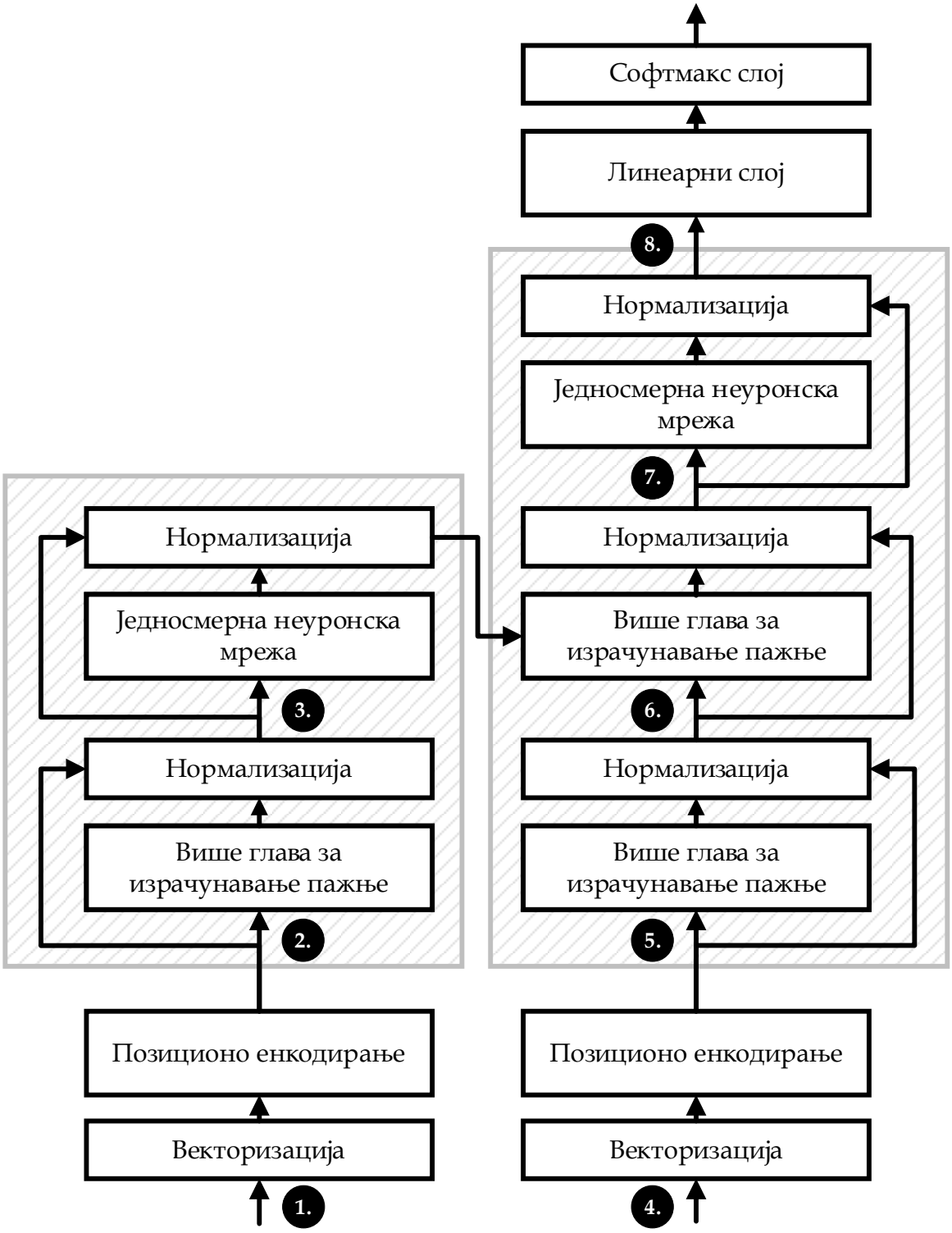
1. Улаз се најпре векторизује користећи предобучени модел за векторизацију токена (обично на основу семантике, изведене из контекста у којем се појављују). Дакле, ниске текста се представљају помоћу вектора бројева, при чему су ниске сличног значења позициониране блиско у векторском простору. Тако векторизован улаз се потом допуњује информацијом о позицијама токена у тексту или сегменту који се обрађује, такозваним, *позиционим кодирањем*, при чему се вредности вектора прилагођавају томе да у векторском простору буду приближније и ниске које се налазе у непосредној близини једна другој у тексту који се посматра;
2. Израчунати вектори се потом прослеђују одређеном броју *глава* (*attention heads*), неуронским слојевима који се обучавају да распознају битне од

небитних токена. Свака глава *мисли за себе* и производи векторе тежине (значајности) токена на основу њихове међусобне повезаности, такозване *маске пажње* (*attention masks*), које се потом множе са оригиналним векторима како би се произвеле нове репрезентације, које се потом нормализују;

3. Добијени вектори се даље прослеђују у једносмерну предобучену ВНМ (због чега је била потребна њихова нормализација), а њен излаз се множи са векторима који су били излаз претходног корака. Добијене векторске репрезентације се поново нормализују, за потребе даље обраде, и тиме се енкодирање завршава;
4. Декодирање започиње на исти начин као и енкодирање. Жељени излаз се векторизује коришћењем истог предобученог модела за векторизацију токена, а резултати се и овде допуњује информацијама позиционог кодирања. Дакле, и овај корак се завршава векторском репрезентацијом резултујуће ниске допуњеном информацијама о позицији токена у том тексту;
5. Добијени вектори се, као и у другом кораку, прослеђују групи глава за израчунавање маски пажње, након чега се вектори множе са израчунатим маскама и резултат се нормализује;
6. Вектори добијени у претходном кораку, заједно са векторима који су настали као резултат енкодера (трећи корак), се прослеђују још једној серији глава, које овога пута сагледавају повезаност између свих токена улаза и жељеног излаза и израчунавају нове маске пажње које се примењују векторе декодера, мењајући их тако да се у њима енкодира и информација о међусобној повезаности са токенима који су послати енкодеру;
7. Вектори добијени у претходном кораку се потом провлаче кроз још једну једносмерну ВНМ и множе са добијеним излазом, како би се добиле нове, измењене репрезентације, које се још једном нормализују, чиме се завршава процес декодирања;
8. Излаз декодера се, коначно, пресликава са свим токенима предефинисаног лексикона токена помоћу још једног линеарног слоја ВНМ. Излаз из тог линеарног слоја представља вероватноће за сваки од токена, које можемо нормализовати примењивањем још једног *софтмакс* (примена нормализоване експоненцијалне функције) слоја у циљу нормализације израчунатих вероватноћа (тако да им сума буде један), што се остварује на следећи начин:

$$\text{Софмакс}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)},$$

где је x_i улазни вектор, а x_j излазни вектор.



Илустрација 11: Архитектура трансформера, где осенчене површине представљају енкодер (лево) и декодер (десно), а бројеви означавају места где започињу кораци.

2.4.2 Врхунски модели

Вероватно најутицајнији језички модел данас је *BERT* (*bidirectional encoder representations from transformers*), *двосмерно кодирање репрезентације из трансформера*, директно заснован на архитектури трансформера, и који су, такође, развили истраживачи компаније Гугл (Devlin, et al., 2018). Оригинално је обучаван на корпусу енглеске Википедије (2.5 милијарде речи) и корпусу *BooksCorpus* (Zhu, et al., 2015) (800 милиона речи), са моделима две различите величине, такозвани $BERT_{BASE}$ (12 енкодера и 12 глава) и $BERT_{LARGE}$ (24 енкодера и 16 глава), који су по објављивању остварили рекордне резултате на великом броју задатака у обради природног језика.

Његова успешност је утицала и на то да се појаве бројне деривације у року од само годину дана, при чему се углавном ради о деривацијама које производе боље резултате у специјалним околностима или направљеним за специфичне домене или употребе:

- *XLM – crosslingual language model* (међујезички језички модел) (Conneau & Lample, 2019) је модел обучаван на текстовима писаним на преко сто светских језика, а са идејом да се може користити за обраду било којег од тих језика, као и да буде оптималан за обраду вишејезичких текстова или текстова за које немамо конкретне информације о језичком пореклу. За сличне потребе може користити и *UDify* модел (Kondratyuk & Straka, 2019), који је такође обучен на вишејезичном, али мањем скупу података.
- *ERNIE – Enhanced Representation through Knowledge Integration (*побољшана репрезентација помоћу интегрисања знања*)* (Sun, et al., 2019) је модел који су развили истраживачи компаније *Баиду*, и који је намењен да се користи у опште сврхе. Унапређење у односу на стандардни модел се проналази у проширењу лексикона фразама и именованим ентитетима.
- *MT-DNN – Multi-Task Deep Neural Networks (*дубока неуронска мрежа за више задатака*)* (Liu, et al., 2019), је још један општи модел, направљен тако да се може користити за обављање више различитих задатака, али је највећу примену нашао у задацима разумевања природног језика, попут препознавања повезаности реченица или њихових контрадикција.
- *ERNIE – Enhanced Language Representation with Informative Entities* (побољшано представљање језика информативним ентитетима) (Zhang, et al., 2019), је модел заснован на интеграцији графова знања у *BERT* архитектуру, и не треба га мешати са претходно поменутим моделом истог акронима. Обучаван је за рад са задацима где се примењују базе знања, али тако да остале задатке обавља на истом нивоу као основни *BERT* модел.
- *MASS – Masked Sequence to Sequence Pre-training for Language Generation* (предобучавање маскираног низа у низ за генерисање језика)*,* (Song, et al., 2019) је модел намењен за генерисање језика, најчешће да допуни део реченице тамо где он фали, а може се користити и у сврху генерисања одговора на питања, сажимања текстова и машинског превођења.

- *XLNet* (Yang, et al., 2019) је једна од најпопуларнијих деривација архитектуре *BERT.* Модел је обучаван је за општу намену, али на знатно већем корпусу у односу на основни модел, па стога очекивано постиже и боље резултате приликом евалуације на преко двадесет разнородних задатака у обради природног језика.
- *RoBERTa – Robustly Optimized BERT Pretraining Approach* (робусно-оптимизовани приступ предобучавању *BERT* модела) (Liu, et al., 2019) је још једна од најпопуларнијих деривација, намењен у општу сврху, а модел је предобучаван још дуже од претходног модела и на још већем корпусу текстова. Услед тога постиже боље резултате од свих модела за општу намену и то на свим задацима. Ипак, како би се смањило време обучавања, овај приступ донекле одступа од предвиђања следеће реченице, што је основни приступ обучавања оригиналног модела.
- *DistilBERT – дестиловани BERT* (Sanh, et al., 2019), у коме су мање значајни параметри пречишћени, има потпуно другачији приступ од претходна два. Користећи методе *преношења знања* (*transfer learning*), омогућава лак начин да се добије модел који је мањи, а самим тим и бржи, и који се може лакше прилагођавати специфичним језицима или наменама.
- *ViLBERT – визиолингвистички BERT* (Lu, et al., 2019) и *VideoBERT* (Sun, et al., 2019) су деривације специјално дизајниране као потпора заједничкој обради текста и слика (први) и текста и видео-снимака (други). Ово се постиже обучавањем векторских репрезентација слика и снимака, те њиховим процесирањем заједно са текстом приликом обучавања модела. Поред тога, *VideoBERT* се може користити и за генерисање видео-материјала на основу унесеног текста и генерисање текста на основу задатог видео-материјала у својеврсној форми машинског превођења између ова два различита медија.
- *UniLM – unified language model* (унификовани језички модел) (Dong, et al., 2019) развили су истраживачи компаније Мајкрософт (*Microsoft*) са циљем да омогући рад на два задатка: генерисању и разумевању природног језика. Попут, *MASS* модела, фокусира се превасходно на задатке сажимања текстова (*text summerization*) и одговарања на постављена питања (*question-answering*).
- *SpanBERT* (Joshi, et al., 2020), најновији модел на овој листи, користи сличан приступ обучавању као *RoBERTa*, јер се не ослања на предвиђање следеће реченице. Његов јединствени допринос је то што се фокусира на боље представљање већих текстуалних целина, комбинујући токене у н-граме приликом обучавања, као и приликом извршавања задатака. Показано је да је овакав приступ бољи за неке од задатака, док у другим показују слабије перформансе од основног *BERT* модела.

Упркос томе што је оригинални *BERT* обучаван на задатку предикције наредне реченице (у односу на контекст, а у затвореном домену реченица), он (за разлику од његових горе поменутих деривација, модела *MASS* и *UniLM*) нема могућност генерисања нових реченица, те га у контексту нашег рада не можемо сматрати псеудограматиком.

2.4.3 Генеративни предобучени трансформери

Паралелно са развојем трансформера *BERT*, истраживачи из *OpenAI* групе, радили су на развоју сопственог трансформера, специфично дизајнираног за генерисање текста, названог *ГПТ* – генеративни предобучени трансформер (*generative pre-trained transformer*). Првобитни модел (Radford, et al., 2018) обучаван је на разноврсном корпусу дугих текстова, у циљу стварања језичког модела који има могућност да повезује чак и појмове који се налазе на великим раздаљинама у корпусу. Показано је, између осталог, да овако обучаван трансформер поседује могућност усвајања широког спектра знања, попут повезивања имена и локација или људи и људских занимања, користећи само неанотирани текстуални корпус као референцу.

Друга итерација из ове групе, *ГПТ-2* (Radford, et al., 2019), настала је убрзо потом, као серија модела различитих величина и могућности објављених у више наврата, услед забринутости да би се могли злоупотребити (за писање лажних вести, електронских порука итд.), јер су на основу кратких упита генерисали смислен текст, наочиглед налик на људски. Модели су обучавани на веб-корпусу од преко осам милиона текстуалних докумената (величине приближне 40GB) прикупљених гребањем страница чије су везе постављене на веб-форуму Редит (*Reddit*), искључујући линкове који воде на Википедију (*Wikipedia*). Објављена су два званична модела која су се разликовала у величини мреже која се обучава: мањи модел је имао око 774 милиона параметара, а већи је имао 1558 милиона параметара (дупло већи од мањег модела). Треба напоменути да је уз то објављено и још неколицина незваничних модела различитих величина од којих је најпопуларнији (поготово у време када званични модели још нису били јавно доступни) такозвани *OpenGPT-2,* отворени ГПТ-2 модел (Cohen & Gokaslan, 2020).

Трећа итерација генеративних трансформера, *ГПТ-3* (Brown, et al., 2020) ставља фокус на дужину обучавања, величину модела и величину корпуса за обучавање, а има за циљ да покаже да су претходни модели ипак били недовољно обучавани и да је више *и даље* једнако боље. Модел који је обучаван за ту прилику има 175 милијарди параметара, те је за цео ред величина већи од претходника. Уз то, и корпус на коме је обучаван је повећан преко пет пута, тако да броји преко 500 милијарди токена, од којих је поново, велика већина настала машинским прикупљањем садржаја са интернета. Убрзо по објављивању истраживања, компанија Мајкрософт је откупила права на ГПТ-3 и има ексклузиван приступ његовом изворном коду, што је утицало и на појаву отворених модела који га емулирају (али са мањим бројем параметара) попут ГПТ-нео (Gao, et al., 2020) и ГПТ-ј (Wang & Komatsuzaki, 2021). Поређења неких од поменутих модела на задатку генерисања текста над четири различита корпуса реченица за тестирање: *LAMBADA* (Paperno, et al., 2016), *Wingrande* (Sakaguchi, et al., 2021), *Hellaswag* (Zellers, et al., 2019) и *PIQA* (Bisk, et al., 2020), приказана су у наставку (Табела 1) (Brown, et al., 2020; Shen, 2022), а задатак на којем је вршена евалуација је исправно довршавање реченица ових корпуса.

Табела 1: Поређење перформанси (тачност погођених наставака реченица) различитих модела заснованих на ГПТ архитектури на задатку генерисања текста над четири различита специјализована корпуса.

Модел	***LAMBADA***	***Winogrande***	***Hellaswag***	***PIQA***
GPT-2 1.5B	51.2%	59.4%	50.9%	70.8%
GPT-Neo 1.3B	57.2%	55.0%	48.9%	71.1.%
GPT-Neo 2.7B	62.2%	56.5%	55.8 %	73.0 %
GPT-J 6B	69.7%	65.3%	66.1%	76.5%
GPT-3	**76.0%**	**88.3%**	**78.1%**	**80.5%**

Генеративни предобучени трансформери имају двоструку употребу: могу евалуирати вероватноћу неке ниске текста или погађати наредни токен за неки задати леви контекст. Ове могућности, о којима ће бити више речи у наредном поглављу их чине савршеним панданом формалне граматике.

3
О генерисању квалитетног текста

Идеја машинског генерисања текста на природном језику најпре се јавила као одговор на потребу за аутоматским конструисањем корисничких интерфејса у рачунарским системима. Убрзо потом су запажене и друге, додатне могућности које су га учиниле једним од циљева истраживача вештачке интелигенције (Mykowiecka, 1991), а олакшавање комуникације са рачунаром се потом проширило на олакшавање комуникације између било које две јединке које *не говоре истим језиком*. Примена генерисања текста на природном језику грубо је подељена у пет група:

- машинско превођење;
- интелигентни едукативни системи (првенствено они за учење језика);
- аутоматско одговарање на питања корисника;
- описивање познатих података природним језиком;
- тестирање теоретских хипотеза у лингвистици и обради природних језика.

Идеја разговора на већ добро познатом, природном језику, док машина врши превођење за нас, упркос примамљивости, није до данашњег дана на задовољавајући начин спроведена у дело.

3.1 Формалне граматике и генерисање текста

Претходно је дефинисано (одељак 1.1) да граматика неког језика генерише све његове реченице. Под условом да је граматика којом рачунарски систем влада беспрекорна, што уобичајено није случај, генерисање природног језика се своди на одабир праве реченице за праву прилику. Ова генерална идеја је оспорива из два разлога. Први разлог је да до сада није направљена беспрекорна граматика природног језика (као језика опште намене, ЈОН), а други да је генерисање бесконачног броја реченица (колико их има у природном језику) немогуће. Неопходан је одабир тј. генерисање праве реченице без провере и генерисања свих осталих кандидата. Свођење

граматике на једну реченицу, може се извршити на два начина: *редукцијом* и *спецификацијом* (Csuhaj-Varjú, 1994).

Метод редукције се примењује тако што се скуп терминала и правила неке граматике смањује, како би та граматика генерисала мањи број реченица. Пример би била редукција енглеског језика на раније приказан опитни подскуп (Пример 1), где је лексикон сведен на три речи и постоје само четири генеративна правила. Такође, како је овај подскуп произведен у циљу специфичне употребе он се може класификовати и као језик посебне намене (ЈПН).

Супротно од њега, метод спецификације има за циљ да граматику учини толико комплексном да она може генерисати мањи, коначан број реченица, што се постиже додавањем и изменом правила као што је показано кроз Пример 2 (са тим да се радило о спецификацији врло једноставне, већ редуковане, граматике). Спецификација се може извршити и смањивањем употребе нетерминала или њиховом заменом одређеним терминалима у новим правилима. Ипак, због велике комплексности (поготово када се ради о спецификацији ЈОН), овај метод се ретко користи.

3.2 Формалне граматике и евалуација генерисаног текста

Други важан аспекат генерисања текста на природном језику јесте његова евалуација. Иако се можемо ослонити на то да ће спецификација или редукција смањити број потенцијалних кандидата, и даље морамо одабрати правог међу њима, што се најбоље види на примеру машинског превођења. Савремени системи који користе велике количине статистичких података, немају проблем да ограниче лексикон и правила те често обављају добар посао на мањим реченицама, али не и на већим, комплекснијим (Пример 3 и Пример 4).

The man took the book. *Човек је узео књигу.*

Пример 3: Превод кратке реченице са енглеског на српски коришћењем сервиса Гугл преводилац (*Google translate*) – 21.08.2018.

Савремени системи који користе велике количине статистичких података, немају проблем да ограниче лексикон и правила те често обављају добар посао на мањим реченицама, али не и на већим, комплекснијим (примери 3 и 4). *Modern systems that use large amounts of statistical data have no problem limiting lexicon and rules and often doing good work on smaller sentences, but not on larger, more complex (Example 3 and 4).*

Пример 4: Пример превода комплексне реченице са српског на енглески коришћењем сервиса Гугл преводилац (Google translate) – 21.08.2018.

Говорник језика који чита овај превод (Пример 4) ће приметити да нешто није у реду, и то је управо оно што оваквим системима највише недостаје – систем евалуације који

ће им рећи да сигурно постоји бољи кандидат од предложеног. Наравно, врхунски системи попут овог се константно побољшавају, што се види на тестирању новије верзије преводиоца на истој реченици нешто више од четири године касније, при чему се добија скоро коректан превод реченице са енглеског на српски језик (Пример 5).

Modern systems that use large amounts of statistical data have no problem limiting the lexicon and rules and often do a good job on smaller sentences, but not on larger, more complex ones (examples 3 and 4)

Пример 5: Пример превода исте комплексне реченице (Пример 4), са српског на енглески коришћењем сервиса Гугл преводилац (*Google translate*) – 20.1.2023.

Евалуација генерисаног текста је у тесној вези са препознавањем језика. Уколико граматика G не може да генерише реченицу која је предмет евалуације, онда та реченица не припада језику L(G) и сигурно није исправна, и самим тим ни прави кандидат. Међутим, реченица не мора да буде прави кандидат ни ако граматика G може да је генерише, с обзиром да она може да генерише све реченице језика L(G)**.** Дакле, погрешан избор кандидата за реченицу генерисаног текста је последица или неисправне граматике G или њене спецификације коришћене при генерисању, као у наведеним примерима машинског превођења (Пример 4 и Пример 5). Тада се морају извршити друге анализе како би се утврдило који је прави кандидат у датој ситуацији.

3.3 Класичне методе оцењивања квалитета текста

Евалуација квалитета текста је без сумње тежак задатак, јер, шта је уопште квалитетан текст? Класичне методе оцењивања попут евалуације корисника, експерта или упоређивањем са *златним стандардом* (еталоном који се сматра идеалним и у односу на који се врши евалуација) теже да буду субјективне, али адекватна алтернатива још увек не постоји. Оцењивање *квалитета* надражаја, било да је у питању песма, филм, или писана реч, мора бити субјективно, јер га различите особе другачије опажају. Различите метрике коришћене при евалуацији покушавају у најбољу руку да *погоде* како би на тај надражај реаговала већина што се мора узети са резервом.

Квалитет текста има неколико особина које нам омогућују да га приближније схватимо као особеност, али и као појам. Он је *некомпозициона* особина текста. Његова граматичност може да игра важну улогу, али исто тако потврда граматичке исправности не мора да значи потврду квалитета. *Епифеноменалан* је , што значи да се не може описати или утврдити без самог примера текста. Квалитет текста је такође под великим утицајем контекста у коме се текст налази. Иста ниска карактера може бити квалитетна у једном и неквалитетна у другом контексту. Такође, разликујемо и пет најчешћих метода оцене квалитета генерисаног текста: експертску евалуацију, упоређивање са златним стандардом, евалуацију на основу учинка по задатку, математички-засноване евалуационе метрике и Тјурингов тест (Hardcastle & Scott, 2008).

3.3.1 Експертска евалуација

Особине: Најстарија метода евалуације текста је она од стране експерта. Идеја је да би доменски експерт могао да распозна квалитетне и неквалитетне текстове у својој области под претпоставком да има искуства са њима јер их чита и/или пише. Овај, врло једноставан концепт заснива се на хипотези да из искуства следи знање. Како смо већ рекли да је евалуација надражаја најчешће субјективна, ова метода је додатно подупрта идејом да је оно што заправо тражимо у тексту квалитет на статистички значајном нивоу. Из истог разлога пожељно је и да се приликом евалуације консултује више експерата.

Недостаци: При тестирању експертске евалуације над кратким текстовима из области енигматике (Hardcastle & Scott, 2008), аутори истраживања се слажу да су добили важне повратне информације од експерата, али наилазе на проблем који је веома чест код ове методе, а то је да се експерти међусобно не слажу око квалитета различитих компоненти текста.

3.3.2 Упоређивање са златним стандардом

Особине: Други најстарији метод заснива се на пренесеном знању експерата. Текстови који се сматрају златним стандардом су они за који већи број, пожељно већина, експерата сматрају да је квалитетан. Ови референтни текстови у широкој су употреби у домену машинског превођења, а користе се и како би се међусобно упоредиле различите евалуације и метрике.

Недостаци: Ова метода наилази на више проблема. Разрешавање питања који текстови би представљали златни стандард је тешко, а поређење текста са њим је још слабије утврђено у теорији. Различити начини поређења често резултују различитим оценама и због тога је контроверзан као метод доказивања.

3.3.3 Евалуација на основу учинка по задатку

Особине: Ова метода потиче из ергономског домена и заснива се на хипотези да ако нешто добро обавља посао, то онда мора бити добро. За потребе истраживања (Hardcastle & Scott, 2008), аутори су генерисали текстове из домена енигматике (прецизније укрштеница, скандинавки) и тестирали њихов квалитет помоћу система који аутоматски (иако то није пресудно) решава укрштенице. Идеја је била да, уколико систем који решава укрштенице успе да реши њихове аутоматски генерисане укрштенице, онда су те укрштенице обавиле *добар посао*, па самим тим морају бити и квалитетне.

Недостаци: Главни проблем ове методе је то што није лако, а често ни могуће пронаћи задатак кроз чије би обављање један систем за генерисање оправдао свој квалитет. Дефинисање задатка, а потом и тестирање изискује исцрпну количину времена, а одвећ често и осмишљени задатак представља слепу улицу.

3.3.4 Математички-засноване метрике

Особине: Методи који користе математичке метрике развијени су у циљу да се при евалуацији избегне проблем субјективности. *Читљивост* текста (упросечена количина прочитаног текста у јединици времена), *граматичност* (обрнута пропорционалност количини граматичких грешака у количини текста) и *флуентност* (број и дужина познатих н-грама у количини текста) само су нека од популарних метрика. У зависности од укључености људске интервенције методи засновани на метрикама се деле се на ручне, полуаутоматске и аутоматске. По правилу, аутоматске методе су брже и јефтиније, али ипак захтевају проверу у виду поређења са ручним, како би се њихова учинковитост дефинитивно утврдила.

Недостаци: Један од проблема коришћења поменутих метрика у евалуацији квалитета јесте њихово везивање за доменски модел. Уколико су текстови нечитљиви и/или неграматични свакако ће се праведно осудити као неквалитетни, али то што су читљиви и граматични не мора бити пресудно за оцену њиховог квалитета.

3.3.5 Тјурингов тест

Особине: Као решење које не подлеже проблемима четири претходно наведена метода предлаже се евалуација у стилу *Тјурингового теста* (Turing, 1950). У раду (Hardcastle & Scott, 2008) аутори су припремили тридесет парова енигматских асоцијација које упућују на исти појам, а анкетираним испитаницима је дат задатак да препознају који од њих је генерисао човек, а који машина, пре него што им је приказано тачно решење. Испитаници су исправно погодили у 72% случајева, а хипотеза је да је то у блиској вези са квалитетом текста асоцијације.

Недостаци: Као негативну компоненту теста аутори ипак наводе да се и он може окарактерисати као субјективан и да је могуће да одговор испитаника зависи од њиховог мишљења или предрасуда које гаје према машинама и машински генерисаном тексту. Ипак, како је полазна основа аутоматског генерисања текста да текст наличи на људски , ову методу тестирања предлажу као најпогоднију.

3.4 Популарне метрике за аутоматску евалуацију генерисаног текста

У циљу стандардизације и лакшег поређења ефикасности језичких модела предложене су, развијене, и користе се различите математички-засноване технике за евалуацију текста. Неке од њих се користе само за евалуацију машински-генерисаног, неке се могу користити и за евалуацију људски-креираног текста, док друге налазе употребу само у специфичним случајевима попут машинског превођења или тестирања новокомпонованих језичких модела.

Према неопходним ресурсима, а за потребе овог рада, поделићемо ове метрике на две групе: *екстринсичне,* спољашње метрике, засноване на задацима, и на евалуацији модела у обављању специфичног задатка (на пример машинског превођења) и на *интринсичне,* унутрашње метрике, засноване на предобученим моделима и независне од задатка за који се ти модели употребљавају, које, дакле, захтевају предобучаване моделе као златни стандард за евалуацију било текста било ефикасности нових језичких модела. Како је број метрика обилан и нове се појављују свакога дана, овде ћемо навести само неке од тренутно најзаступљенијих, или карактеристичних за методе битне за овај рад.

3.4.1 Екстринсичне метрике

Ове једноставне метрике, обично засноване на н-грамској анализи, користе се за евалуацију машински генерисаног текста, не према језичком моделу који генерише текст, већ према задатку за који је модел развијен (углавном задаци машинског превођења или препознавања говора). Екстринсичне метрике се заснивају на претходно обележеним подацима, на пример, уређеним паровима текстова и њихових ручних превода (уколико се ради о евалуацији машинског превођења), уређеним паровима звучних и текстуалних записа говора (уколико се ради о препознавању говора) итд. Машински генерисан текст се на различите начине пореди са референтним текстом и додељује му се нека нумеричка вредност која означава његову репрезентативност на различите начине, од којих ће неки бити описани у наставку текста.

WER (*Word error rate*) – *постотак погрешних речи* (Klakow & Peters, 2002) је најосновнија метрика за евалуацију машински генерисаног текста, а користи једноставну формулу засновану на Левенштајновом растојању (Левенштейн, 1965):

$$WER = \frac{S + D + I}{N}$$

где је N укупан број речи у тексту који се евалуира, S је број неопходних замена речи, D број неопходних брисања, а I број неопходних додавања. Дакле, *WER* је једнак количнику укупног броја грешака које треба исправити и укупног броја речи.

TER (*Translation error rate*), *постотак погрешног превода* (Agarwal & Lavie, 2008), је унапређена верзија *WER* метрике, такође заснована на Левенштајновом растојању, али која узима у обзир додатне могућности приликом исправљања грешака, попут померања речи удесно и улево за једно или више места.

BLEU (*bilingual evaluation understudy*) — двојезична евалуативна метрика (Papineni, et al., 2002) је тренутно најпопуларнија екстринсична метрика за евалуацију, а користи нешто компликованији алгоритам за израчунавање репрезентативности јер, за разлику од *WER*, није заснована на речима, већ на њиховим н-грамима (где се распон *н* подешава арбитрарно, а обично је у распону од један до четири). За свако задато *н* израчунава се н-грамска прецизност (постотак референтних n-грама у генерисаном тексту), потом се израчунава геометријска средина свих добијених н-грамских прецизности. Такође се примењује и казнена стопа за генерисане реченице које су краће од референтних репера, како би се избегло увећање коначног резултата услед малог делиоца при израчунавању прецизности. Дакле *BLEU* се израчунава на следећи начин:

$$BLEU(N) = B \times P(N)$$

где *N* дефинише степен н-грама, *B* је казнена стопа краткоће (*brevity penalty*), а *P(N)* геометријска средина н-грамских прецизности (p_n) која се израчунава као:

$$P(N) = \exp\left(\sum_{n=1}^{N} \frac{1}{N} \ln p_n\right)$$

где је функција $\exp(x)$ једнака e^x.Казнена стопа краткоће *B* израчунава се на следећи начин:

$$B = \begin{cases} 1, r < c \\ e^{\left(1-\frac{r}{c}\right)}, r \geq c \end{cases}$$

где је *r* број речи у генерисаном тексту, а *c* број речи у референтном тексту.

Метрика *NIST* (*National Institute of Standards and Technology*) је метрика Националног института стандардизације и технологије Сједињених Америчких Држава (Doddington, 2002), настала као унапређење и уопштење метрике *BLEU*. Главна разлика између тих метрика је то што се, уз прецизност, израчунава и информативност сваког н-грама тј. мање фреквентни н-грами носе већу тежину и више утичу на укупни резултат.

$$NIST(N) = B \times PP(N)$$

где је *PP(N)* пондерисана (тежинска) геометријска средина н-грамских прецизности подешених тако да узимају у обзир и информативне тежине свих н-грама (w_n):

$$PP(N) = \exp(\sum_{n=1}^{N} w_n \ln p_n)$$

при чему су w_n позитивни бројеви чији је збир 1. Ако су све тежине w_n међусобно једнаке, метрика *NIST* се своди на *BLEU*. Казнена стопа краткоће *B* израчунава се на исти начин.

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) је скуп одзивно-оријентисаних метрика за евалуацију (Lin, 2004), који се, за разлику од претходно наведених метрика, користи превасходно за евалуацију машински-генерисаних сажетака текста, и није фокусиран на н-грамску прецизност, већ на н-грамски одзив, а рачуна се као:

$$ROUGE(N) = B \times R(N)$$

где је *R(N)* геометријска средина н-грамских одзива.

METEOR (*Metric for Evaluation of Translation with Explicit Ordering*)—метрика за евалуацију превода са експлицитним редоследом (Banerjee & Lavie, 2005) је још једна варијанта *BLEU* метрике на коју је утицала и појава *ROUGE*, а заснива се и на униграмској прецизности и на униграмском одзиву, стављајући посебан акценат на однос одзива и прецизности тј. на F-меру:

$$METEOR(N) = F(1 - p)$$

где *F* представља Ф-меру (хармонијску средину прецизности и одзива) добијену на основу униграмских прецизности и одзива, а p је коефицијент који служи да умањи резултат тамо где редослед речи није у складу са референтним текстом.

3.4.2 Интринсичне метрике

Интринсичне метрике се не ослањају директно на задатке ни референтне текстове, већ се за златни стандард узимају модели претходно обучавани на референтним текстовима. Уобичајена пракса је да се два модела пореде и да се анализира колика је вероватноћа да ће генерисати исти текст.

Најчешћи облик интринсичне евалуације у рачунарској лингвистици остварује се коришћењем *перплексности*, мере *изненађености* модела неким текстом. Перплексност представља реципрочну вредност вероватноће да неки модел генерише неку ниску токена (тј. текст), нормализовану дужином тог текста (Brown, et al., 1992). За неку ниску токена, при чему вероватноћа зависи и од контекста (ниске токена који претходе), ентропија се рачуна као просечна количина информација коју доноси нова реч. У складу са тиме, перплексност (*PP*) се у служби мере квалитета језичког модела може израчунати на следећи начин:

$$PP = \left(P(w_1 w_2 \dots w_n)\right)^{-\frac{1}{n}}$$

где је $(w_1 w_2 \dots w_n)$ ниска токена, P њихова вероватноћа, а n дужина ниске, што се може написати и као:

$$PP = \sqrt[n]{\frac{1}{P(w_1 w_2 \dots w_n)}}$$

Уколико је вероватноћа реализације неке ниске токена према одређеном моделу једнака *0.25*, а дужина ниске је *2*, онда ће перплексност бити израчуната на следећи начин:

$$PP = \sqrt[2]{\frac{1}{\mathbf{0.25}}}$$

$$PP = \sqrt[2]{4}$$

$$PP = 2$$

Дакле, што је вредност перплексности већа, задата ниска слабије одговара моделу који се тестира, или модел слабије одговара тексту на ком се тестира.

Са тим на уму, уколико имамо текст који је златни стандард, перплексност можемо користити као меру квалитета модела, или можемо мерити квалитет генерисаног текста уколико узимамо предобучавани модел за златан стандард који желимо да постигнемо. У оба случаја тражимо да мера перплексности буде што нижа, било да тражимо што вероватнију ниску за неки задат контекст или желимо да тестирамо способност наше псеудограматике да генерише кохерентан, вероватан текст. У најгорем случају, када је модел потпуно неспреман и вероватноћа за сваки токен је иста, онда је перплексност једнака величини алфабета токена тј. лексикона **Σ**.

Осим општих мера попут перплексности, које се могу користити за поређење језичких модела и текстова на неком апстрактном нивоу, постоје и мере засноване на специфичним, моделима, попут *BERTscore* (Zhang, et al., 2019) и *BLEURT* (Sellam, et al., 2020), примењене у системима за одређивање квалитета текста заснованим на *BERT* архитектури и/или *BERT* језичким моделима (одељак 2.4.2), а улазни текст се анализира коришћењем специјалних предобучаваних модела. Метрика *BERTscore* омогућава израчунавање семантичке сличности између две реченице, па је добра пре свега за задатке машинског превођења где су унапред дефинисани уређени парови, као код екстринсичних метрика. Са друге стране, метрика *BLEURT* класификује реченице на основу њиховог квалитета, користећи и дообучавање на скупу реченица које је претходно евалуирао и класификовао човек.

3.5 Хибридна евалуација квалитета текста

Главна разлика између аутоматске и ручне евалуације је у ресурсима који се користе. Док је за ручну евалуацију неопходно време и одговарајућа радна снага, за аутоматску евалуацију неопходни су ресурси у виду припремљеног скупа података, који се најчешће ручно анотира, у циљу повећане тачности. Предност аутоматске евалуације је у томе да се једном припремљен ресурс може користити неограничен број пута без значајног додатног улагања, а самим тим је и значајно бржи од ручног. Ипак, ручна евалуација и данас важи за златни стандард евалуације. У раду (Belz & Reiter, 2006) упоређују се евалуације генерисаног текста, ручна (доменски експерти и испитаници без доменске експертизе) и аутоматска (претходно поменути *NIST*, *BLEU* и *ROUGE* системи). Након упоређивања резултата аутори закључују да аутоматска евалуација има потенцијала, али да мора даље да се развија. Њено потпуно ослањање на статистичке методе у најпопуларнијим системима тог времена и, на изглед, само површинско разумевање, јесте оно што је понајвише разликује од ручне евалуације у којој је анализа комплекснија.

Са тим у виду, (Pitler & Nenkova, 2008) предлажу ново, хибридно окружење за предвиђање квалитета текста. Аутори аутоматски израчунате статистичке податке комбинују са ручно добијеним, експертским оцењивањем квалитета текста у три категорије: лексика, синтакса и дискурс. Резултати комбиновања показују да оцене квалитета експерата из домена лингвистике побољшавају резултате детекције квалитетног текста за приближно 6% у односну на чисти, статистички метод. Како овај метод захтева мање ресурса од ручног, а више од аутоматског, он производи и резултате сходне томе (боље од аутоматског али лошије од ручног метода).

Не искључује се, међутим, могућност да би лингвистичка обележја могла бити и аутоматски прикупљена из ресурса, чиме би се цена овог метода значајно снизила, а било би теоретски могуће добити подједнако добре резултате. Експерименти са екстракцијом оваквих и сличних података из текста значајно су олакшани свеопштом експлозијом количине текстуалних корпуса који су доступни на интернету.

3.6 Алтернативне методе у евалуацији генерисаног текста

Наравно, евалуација генерисаног текста се не мора нужно обављати методама наведеним у одељцима 3.2–3.5. Истраживачи са Универзитета у Сао Паолу (Antiqueira, et al., 2007) преузимају нешто другачији приступ и успостављају директну корелацију између одлика графа комплексне мреже (топологије и тежинских вредности) и квалитета генерисаног текста. Њихов експеримент се састојао од двоструке упоредне анализе есејских текстова ученика средњих школа, при чему су квалитет текстова најпре евалуирали доменски експерти, а ти резултати су потом поређени са онима добијеним трансформацијом текста у комплексну мрежу, анализом њене топологије и аутоматском евалуацијом.

Трансформација лематизованог текста, из ког су претходно уклоњене стоп-речи, у комплексну мрежу обављена је на два начина, тако да резултати трансформације представљају *Марковљеве моделе* са меморијом 1 и 2, а обе мреже су тестиране на корелацију резултата. Прва мрежа, *НЕТ-А,* добијена је претварањем свих речи текста у чворове, након чега су између узастопних речи (на удаљености један) успостављене тежинске везе у зависности од фреквенције њиховог заједничког појављивања (слично анализи друштвених мрежа). Друга мрежа, *НЕТ-Б,* је направљена на исти начин, при чему су додате и везе између речи на удаљености два. Из припремљених комплексних мрежа је за потребе експеримента екстраховано пет атрибута који су коришћени за евалуацију:

1. *Улазни степен мреже,* УСМ, израчунат је као аритметичка средина суме улазних веза (УВ) свих чворова мреже:
$$\text{УСМ} = \frac{1}{n}\sum_{i=1}^{n} \text{УС}(i)$$
где *i* представља сваки од чворова мреже (којих је укупно *n*), а УС се израчунава као:
$$\text{УС}(i) = \sum_{j=1}^{n} W(j,i)$$
где *j* представља сваки од чворове мреже (чији је укупан број *n*), а $W(j,i)$ представља број веза од *j* ка *i*.
2. *Излазни степен мреже,* ИСМ, израчунат је као аритметичка средина суме излазних веза свих чворова мреже:
$$\text{ИСМ} = \frac{1}{n}\sum_{i=1}^{n} \text{ИС}(i)$$
где *i* представља сваки од чворова мреже (којих је укупно *n*), а ИС се израчунава као:
$$\text{ИС}(i) = \sum_{j=1}^{n} W(i,j)$$
где *j* представља сваки од чворове мреже (чији је укупан број *n*), а $W(i,j)$ представља број веза од *i* ка *j*.
3. *Коефицијент груписања мреже,* КГМ добијен је као аритметичка средина коефицијената груписања за све чворове везе, где је коефицијент груписања неког чвора *i* једнак количнику укупног броја веза између свих чворова повезаних са њим (E_i) и разлике квадрата и броја чворова (N_i) који су са њим повезани:
$$\text{КГМ} = \frac{1}{n}\sum_{i=1}^{n} \frac{E_i}{{N_i}^2 - N_i}$$
Идеја овог атрибута је да се одреди колика је тенденција мреже да се групише, насупрот хетерогеној расподели улазних и излазних степени чворова у мрежи.
4. *Девијација динамике мреже,* ДДМ, добијена је као девијација од линеарне еволуције чворова и веза у мрежи. При линеарном развијању мреже број

одвојених компоненти (у овом случају неповезаних чворова) би уједначено опадао са порастом броја веза, док би свако одступање од тога повећало коефицијент ДДМ.

5. *Најкраћи пут мреже,* НПМ—аритметичка средина најкраћег пута израчунатих помоћу *Флојд-Варшаловог* алгоритма (Floyd, 1962) између свака два чвора, са изузетком пута од сваког чвора до самог себе, једнак је и фитнес функцији *n-медијане* која се користи за решавање проблема расподеле објеката:

$$\text{НПМ} = \frac{1}{n} \sum_{i=1, j=1, i \neq j}^{n} \text{ФВ}(i, j)$$

где су i и j различити чворови мреже, а ФВ(i, j) најкраћи пут између њих израчунат помоћу Флојд-Варшаловог алгоритма.

Након упоређивања ових параметара са оценама људских судија за исти текст пронађена је корелација за ИСМ, КГМ и ДДМ и то пре свега за оцену *кохезије и кохеренције*, које одражавају начин на који су реченице и повезане и њихову међусобну складност. За оцену стила писања су пронађене само *могуће* корелације, док за оцену адекватности теме нису пронађене корелације, пре свега, као последица изузетка семантичке анализе у експерименту. Такође, није пронађена статистички значајна разлика између параметара мрежа *НЕТ-А* и *НЕТ-Б*.

Исти тим научника је касније спровео још један сличан експеримент у коме је показано да се упоређивањем комплексних мрежа, сачињених за два различита текста, може утврдити њихов степен корелације (Antiqueira, et al., 2009). Системи развијени за потребе тог експеримента (укупно петнаест њих) коришћени су за аутоматско сажимање текста, а очигледна мањкавост у семантичкој анализи заобиђена је једноставним поређењем топологија мрежа, изузимајући њихов садржај.

Кандидати за сажимање су креирани аутоматски, и у поређењу са комерцијалним системима за аутоматско сажимање који захтевају велику количину лексичких ресурса показали су незнатно слабије резултате. Ови експерименти су показали употребну моћ трансформације текста у комплексне мреже, пре свега у њиховој могућности да се текстови међусобно пореде на нелинеаран начин и тако се између њих успостави корелација у циљу аутоматске класификације и евалуације текста. Добијени резултати су били на нивоу најбољих у то време у области сажимања докумената.

Истраживачи са Универзитета у Оксфорду полемишу да је управо коректна репрезентација реченица у тексту неопходна за адекватно разумевање и обраду текста на природном језику (Kalchbrenner, et al., 2014). У свом раду они описују још једну алтернативну методу реконструкције текста на природном језику коришћењем *динамичких конволуционих неуронских мрежа,* ДКНМ. Реченице *пролазе* кроз ДКНМ и као резултат су генерисани графови који представљају њихову структуру узимајући у обзир и међусобну повезаност структура унутар реченице. За потребе експеримента трансформисали су у графове више различитих текстова у циљу

њихове класификације овом методом и у три од четири теста остварили су статистички значајну прецизност.

3.7 Корпуси квалитетног текста

Већ са појављивањем феномена *Big Data* средином прве деценије двадесет и првог века, постало је јасно да је неопходно одвојити значајан или квалитетан садржај од неквалитетног како би се он на прави начин искористио. Очигледно је да је одређена количина података неупотребљива и тако истраживање квалитета података и текстова на интернету добија на значају. У раду (Agichtein, et al., 2008) аутори покушавају да у текстуалном корпусу прикупљеном са *Yahoo! Answers* портала направе поделу садржаја на квалитетни и неквалитетни. Коришћењем машинских метода класификације који се ослањају на аутоматски прикупљене атрибуте попут оцене корисника или предефинисаних израза, они успевају да са великом прецизношћу класификују квалитетне и неквалитетне одговоре на порталу. Овакве методе анализе текста, данас су широко заступљене и представљају основу при анализи података на интернету (анализе квалитета, сентимента и друге).

Новија истраживања окрећу се и технологијама за реструктурирање текста јер се текст у облику ниске више не сматра нужно најпогоднијим за обраду. Различите методе његовог уобличавања и структурирања постале су неопходност. У раду (Liu, et al., 2015) аутори експериментишу са корпусом текстова подељеним на токене, који не представљају само речи, већ и фразе. Текстови се затим обележавају као квалитетни или неквалитетни и обрађују методама машинског учења. Експериментом је показано да је анализа квалитета текста коришћењем обележених фраза много успешнија у односу на ону која користи само обележене речи. Ови резултати потврђују претходна истраживања (Langner, 2010) где се помоћу обучавања над фразно-структурираним корпусом квалитетног текста дошло до квалитетног система за генерисање природног језика чији је циљ да имитира људски текст.

Сви претходно поменути системи, као и многи други ипак умногоме зависе од квалитета корпуса над којим се обучавају или над којим се генеришу њихове граматике, те је развој корпуса био и остао један од истраживачких приоритета. Закључујемо да граматике и/или псеудограматике добијене из квалитетних корпуса могу бити од пресудног значаја у аутоматској анализи и евалуацији различитих аспеката теста на природном језику.

4
О предмету и циљевима истраживања

Премда нам теорија формалних језика и аутомата нуди начин за описивање природних језика путем формалних граматика заснованих на правилима (Chomsky, 1956), њихово креирање се у пракси тешко примењује због комплексности припреме и укупног времена потребног за њу. Са убрзаним развојем технологија вештачке интелигенције, а посебно метода машинског учења јавила се идеја да се на бржи начин може доћи до доброг резултата креирањем апроксимација тих граматика, заснованих на вероватноћи, статистици и методама машинског учења. У истраживању (Giles, et al., 1992), након експеримената на неколико различитих корпуса, закључено је да рекурентна вештачка неуронска мрежа обучавана на коначном скупу реченица, одговара аутомату који производи бесконачан број реченица регуларног језика.

Развој области обраде природног језика, пре свега интелигентних, статистички-заснованих језичких модела, нам је у последњих неколико година донео и многобројне нове методе и технологије апроксимације језика (Kalchbrenner, et al., 2014), при чему се из године у годину појављују нови, уверљиво бољи језички модели, фокусирани како на специфичне задатке тако и на опште моделирање језика, а најистакнутији примери су *BERT* (Devlin, et al., 2018) и ГПТ (Radford, et al., 2018; Radford, et al., 2019; Brown, et al., 2020), претходно описани у одељку 2.4.

У наредним одељцима биће више речи о предмету истраживања ове докторске дисертације по тезама, потом конкретним циљевима и истраживачким питањима на које она покушава да одговори. Након тога биће описан ток истраживања и коначно кратак преглед очекиваних научних доприноса.

4.1 Циљеви докторске дисертације

Општи циљеви ове докторске дисертације обухватају развијање методологије креирања композитних интелигентних система за решавање задатка изградње псеудограматике, као и постављање темеља за даље истраживање композитних интелигентних система и њихову примену.

Специфични циљ ове дисертације је развијање потребних *појединачних* језичких модела, као и софтверског решења за имплементацију композитних система (заснованих на њиховој *паралелној* употреби и хеуристикама) који ће моћи да се користе као псеудограматике природног, српског, језика. Да би овај циљ могао бити остварен прво ће бити развијени најсавременији поменути језички модели за српски језик, да би се потом, кроз софтверско решење, које ће бити засновано на њиховој паралелној употреби имплементирале одговарајуће псеудограматике.

Да би то постигла, ова дисертација најпре истражује:

- Делотворност коришћења паралелних структура у обради природног језика (Vogler & Metaxas, 1999; Zhang, 2004), са фокусом на српски језик (детаљније у одељку 5);
- Утицај анотације у циљу добијања различитих репрезентација текста, те компаративну анализу паралелних система који користе или не користе те различите репрезентације (детаљније у одељцима 5 и 6);
- Утицај различитих начина комбиновања излазних вредности самосталних система (попут вероватноћа ознака врста речи или удаљености докумената) како би се кроз паралелну архитектуру произвели оптимални резултати (детаљније у одељку 6);
- Резултате постигнуте тренутно најсавременијим методама моделирања језика, као и моделе засноване на различитим информацијама екстрахованим из текста, а потом и идеју креирања композитних интелигентних система заснованих на тим језичким моделима (Hacioglu & Ward, 2001; Broman & Kurimo, 2005; Arefyev, et al., 2019) и њихову примену на пољу препознавања и обраде природног језика (детаљније у одељцима 7, 8, и 9).

Главни фокус ове дисертације је, дакле, проналажење ефикасне архитектуре за *комбиновање* неколицине *појединачних савремених језичких модела* који се *паралелно* користе, у јединствен *систем заснован на хеуристикама*, *псеудограматику*, са акцентом на могућности надоградње, како би се кроз њега могли искористити не само постојећи, већ и будући језички модели, са свим својим предностима.

Сврсисходност понуђеног решења и употребљивост изграђених *псеудограматика над паралелним језичким моделима* ће у раду бити илустрована на примеру решавања задатака попут евалуације докумената и њихове бинарне класификације, те генерисања текста и евалуације његовог квалитета.

4.2 Истраживачка питања

Основна претпоставка од које се полази при раду на докторској дисертацији је да се на основу пробабилистичких излаза више различитих језичких модела, било заснованих на различитим технологијама или на различитим репрезентацијама текста (нпр.: текст у коме су речи сведене на лему или граматичку категорију) примењених над унапред анотираним текстом, може обучити интелигентни систем чије ће перформансе надмашивати перформансе појединачних модела. За потребе истраживања осмишљено је неколико питања на која ће бити одговорено:

- **ИП1** Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и њиховим излазним вероватноћама адекватна метода у области обраде природних језика?
- **ИП2** Да ли се композитни интелигентни системи засновани на паралелним моделима могу надограђивати – и да ли већи број активних паралелних модела побољшава квалитет целокупног система?
- **ИП3** Да ли паралелни језички модели боље моделирају језик од појединачних?
- **ИП4** Да ли композитне псеудограматике боље моделирају језик од појединачних?
- **ИП5** Који су оптимални методи комбинације излаза појединачних језичких модела?
- **ИП6** Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и излазним вероватноћама подобна метода у области интелигентних система уопште?

4.3 Ток истраживања

Претходно су утврђени појмови језика и граматике, те су представљене хипотезе о природи њиховог настанка и могуће поделе (одељак 1). Након тога, уведени су појмови псеудограматике и језичког модела, и представљене су њихове особине, историјски развој и тренутно стање (одељак 2). Коначно, било је више речи о тексту и текстуалним документима, њиховом генерисању, квалитативним особинама и различитим начинима евалуације и утврђивања њиховог свеукупног квалитета (одељак 3). У овим одељцима преовлађује дескриптивна метода која укључује описивање, прикупљање и систематизацију података, те компаративну и аналитичку истраживачку методу која подразумева упоређивање, вредновање и интерпретацију добијених резултата, као и анализу података из истраживања других аутора.

Са друге стране, у одељцима 5 и 6 биће описани експерименти који су за циљ имали да установе делотворност коришћења паралелних структура у обради природног језика, и то, пре свега, српског, као и да утврде методологију креирања композитних система за ове потребе. Ови одељци ће имати најразнороднију методолошку потпору. За производњу потребних корпуса и репрезентација докумената биће примењене методе корпусне и рачунарске лингвистике у комбинацији са методама вештачке интелигенције. За развој већине композитних система биће примењиване методе

вероватноће и статистике као и методе машинског учења, и то претежно вештачке неуронске мреже.

Одељак 5, који има за циљ да одговори на ИП1 и ИП2 бави се истраживањем паралелног процесирања у обради српског језика тестираног на задатку обележавања врсте речи (Stanković, et al., 2022). Током тог истраживања, а за потребе одговора на ИП1, развијен је композитни систем за аутоматско обележавање корпуса текстова врстом речи и лемом, заснован на излазима појединачних модела и наслаганом класификатору заснованом на хеуристикама. Додатно, у овом одељку је описана и компарација система различитих величина (заснованих на различитом броју појединачних модела) која директно потпомаже пружање одговора на ИП2.

Одељак 6, који има за циљ да анализира ефективност коришћења паралелних језичких модела, и утврди и упореди начине њиховог комбиновања детаљније описује експеримент са моделовањем мини-језика коришћењем паралелних језичких модела, које се евалуира на задатку одређивања ауторства (Škorić, et al., 2022). Приликом решавања овог задатка припремљене су четири различите репрезентације докумената (као базе за моделирање језика), и описано је шест различитих метода комбиновања резултата, од којих је пет засновано на правилима и једна на хеуристикама. Комбинације различитих репрезентација и метода комбиновања дале су седамнаест јединствених резултата који су међусобно поређени у циљу добијања одговора на ИП3.

У одељку 7 биће описано утврђивање корпуса српског језика, као и креирање његових различитих репрезентација кроз анотацију засновану на софтверу описаном у одељку 5. Потом ће бити описано и обучавање савремених језичких модела над тим различитим репрезентацијама, а за постизање постављеног циља, биће примењене најсавременије методе за развој језичких модела, генеративни предобучени трансформери, и све то у комбинацији са савременим софтверским окружењима и решењима. Одељак 8 ће описати комбиновање припремљених модела у јединствене псеудограматике српског језика, засноване на претходно утврђеним методама комбиновања описаним у одељку 6.

Евалуација добијених система на неколико припремљених задатака биће описана у одељку 9, при чему ће се користити претежно методе аутоматске евалуације, а евалуација ће обухватити квантитативну анализу добијених резултата засновану на интринсичним метрикама. Резултати, који ће бити приказани у истом одељку, помоћи ће нам да одговоримо на ИП4, ИП5 и ИП6.

У коришћеним методама, као и у примени спроведеног истраживања, огледа се мултидисциплинарност овог истраживања. Поменуте методе припадају различитим областима науке, а пре свега рачунарства и лингвистике (јер је и сама област обраде природног језика мултидисциплинарна), док велику улогу играју и методе вероватноће и статистике, као и машинског учења. Такође, и развијени ресурси и технологије ће наћи примену и у рачунарству и у лингвистици.

4.4 Очекивани научни доприноси

У дисертацији се разматра проблем моделирања језика, при чему се посебна пажња посвећује питању коришћења већег броја расположивих ресурса. Треба узети у обзир да ресурс није само корпус текстова, на чијем коришћењу се заснивају модерна истраживања моделирања језика, већ и различите технологије моделирања као и други, претходно припремљени језички ресурси, који потпомажу дубље разумевање текста. Имајући у виду растући број ресурса и технологија, јавила се потреба за одговарајућим решењем, у виду композитног модела које ће их објединити.

На основу досадашњих, текућих и планираних истраживања, поред опште анализе проблема, научни доприноси овог истраживања који се очекују су:

- Проширење корпуса савременог српског језика;
- Развијање новог софтвера за аутоматску анотацију корпуса, на основу прикупљених предобележених примера;
- Развијање савремених језичких модела српског језика на основу различитих репрезентација корпуса савременог српског језика;
- Развијање детаљног модела композитног система за паралелно обједињавање креираних модела (укључујући и будуће моделе);
- Креирање псеудограматика српског језика које ће имати примену у задацима обраде природног језика, укључујући класификацију и евалуацију докумената, као и генерисање текста;
- Тестирање и валидација на постојећим, претходно истраженим проблемима.

Псеудограматике развијене током овог истраживања наћи ће примену у решавању различитих проблема у обради природног језика, јер се, попут формалних граматика, и оне се могу користити у морфосинтаксној анализи текста, за генерисање новог текста, као и за прецизније израчунавање сличности између текстова. Прикупљено знање и резултати би такође омогућили развој метода који би се користио за креирање нових, или за обједињавање постојећих интелигентних система.

II Паралелни језички модели српског језика

5
Паралелно процесирање у обради српског језика

5.1 Обележавање врстом речи

Обележавање (*тагирање*) токена (или група токена) врстом речи (именице, глаголи, придеви итд.), је добро познат и један је од најуобичајенијих задатака обраде природног језика (ОПЈ), а може се рећи и да је то најосновнији и најчешће коришћен облик анотације текста. Налази примену у многим сферама ОПЈ, попут класификације докумената, препознавања именованих ентитета, анализе сентимената и одговарања на питања (Abney, 1997). Софтвер који обавља задатак обележавања обично се назива *тагер* и може бити базиран на правилима (користећи табеле претраживања, речнике и/или екстрахована лингвистичка правила), стохастици (користећи различите методе машинског учења над обележеним текстом) или хибридни, комбинујући та два приступа. На пример, хибридни тагер би био *TreeTagger*, софтвер за тагирање који користи и табеле за претраживање и речнике (предобележене листе токена и њихових могућих врста речи) заједно са стохастичким приступом *скривених Марковљевих модела – HMM* (Schmid, 1999).

Разлог због којег причамо о обележавању врстом речи (у даљем тексту тагирање) је то што је оно добар медијум за тестирање нових приступа у ОПЈ (и класификације уопштено), превасходно услед своје распрострањености и једноставности. Сваки пут када се јави нови популаран напредак (у методолошком смислу), на пример, нови метод машинског учења, убрзо се појави и нови тагер врстом речи (у даљем тексту тагер) који користи ту нову методу. Из тог разлога данас имамо велики број различитих тагера, и из тог разлога је најпре на задатку тагирања тестиран (Stanković, et al., 2022) и мулти-модални приступ који ће се користити у овој дисертацији, са циљем да се увиди да ли је он погодан за обраду природног (српског) језика.

5.2 *Паралелно тагирање*

Један од најутицајнијих експеримената који се тиче комбиновања више тагера у паралелну структуру тиче се тагирања шведског језика (Sjöbergh, 2003), где је на основу евалуације закључено да већина композитно-заснованих тагера даје боље резултате од најбољег самосталног тагера. Други значајан покушај урађен је за исландски језик (Henrich, et al., 2009), где је побољшање тачности од скоро два процента добијено коришћењем једноставних алгоритама за комбиновање (што је велики успех на задатку тагирања врстом речи, за који су стопе грешке иначе врло мале).

Ови радови су успоставили основни скуп техника комбиновања појединачних тагера у композитну структуру:

- *Гласање* је техника комбиновања где се за коначан одговор (у виду јединствене ознаке за један токен) бира ознака са највише гласова од стране различитих тагера. Главни проблем са овим приступом је у томе што он уопште не функционише када су доступна само два тагера, а чак и их ако има, постоји велика шанса да дође до нерешеног резултата при гласању, што се не може разрешити без имплементације додатног алгоритма.
- *Пондерисано гласање* нуди решење претходно поменутог проблема у виду примењивања тежина на различите ознаке или тагере при гласању, што смањује шансу нерешеног резултата.
- *Бодовање* је нешто комплекснији метод који захтева да тагери ознакама додељују вероватноће. Комбиновање резултата се састоји од сабирања (или још боље, усредњавања) вероватноћа по ознакама и по тагерима, те се ознака која има највише бодова тј. највећу вероватноћу узима за коначан одговор.
- *Лицитирање* је засновано на сличном принципу као бодовање и такође захтева вероватноће, али се уместо сабирања или усредњавања за коначну узима ознака са највећом вероватноћом међу свим ознакама и свим тагерима.

Ипак, најновија прекретница у развоју композитних тагера је техника *слагања*. Овај приступ се користи како би се елиминисао главни узрок ниских перформанси осталих композитних тагера, а то је самостални тагер, са ниским учинком, који нуди велике вероватноће за свој избор. Оно се заснива на додавању још једног *наслаганог* класификатора поврх резултата самосталних тагера те, ако је тај класификатор добро обучен, он зна да се не може веровати тагеру са лошим учинком. Иницијално истраживање излазних вероватноћа самосталних тагера као карактеристика за наслагани класификатор (Aliwy, 2015) произвело је композитни систем, који користи вероватноће (за сваку ознаку и сваки самостални тагер) произведене од неколико самосталних тагера као карактеристике за ХММ класификатор, који производи коначну ознаку за сваки токен.

5.3 Обучавање паралелних тагера врстом речи за српски језик

5.3.1 Анотирани ресурси

Обучавање и тестирање свих тагера за овај експеримент обављено је коришћењем јавно доступног анотираног корпуса српског језика, *SrpKor4Tagging*[1] (342.804 обележена токена, од којих су 306.352 речи) (Stanković, et al., 2020). Отприлике трећина токена овог корпуса потиче из књижевних текстова (романа и одломака романа), док остатак потиче из некњижевних текстова (новински чланци, уџбеници и административни текстови). Детаљи о документима који сачињавају овај корпус (укупан број токена, речи и број јединствених речи) приказани су у наставку (Табела 2). Корпус је помоћу *Unitex* система (Paumier, et al., 2009) подељен на реченице и предобележен коришћењем два скупа ознака: *Universal POS*[2] (седамнаест ознака) и *СрпЛемКор*[3] (шеснаест ознака, развијен за морфолошке речнике српског језика, *СМД* (Krstev, 2008; Vitas & Krstev, 2012), у складу са традиционалном, описном српском граматиком), а та анотација је потом проверена и исправљена ручно (тамо где је то било потребно).

Табела 2: Документи који сачињавају *SrpKor4Tagging* анотирани корпус, који је послужио за израду тагера приказаних у овој секцији.

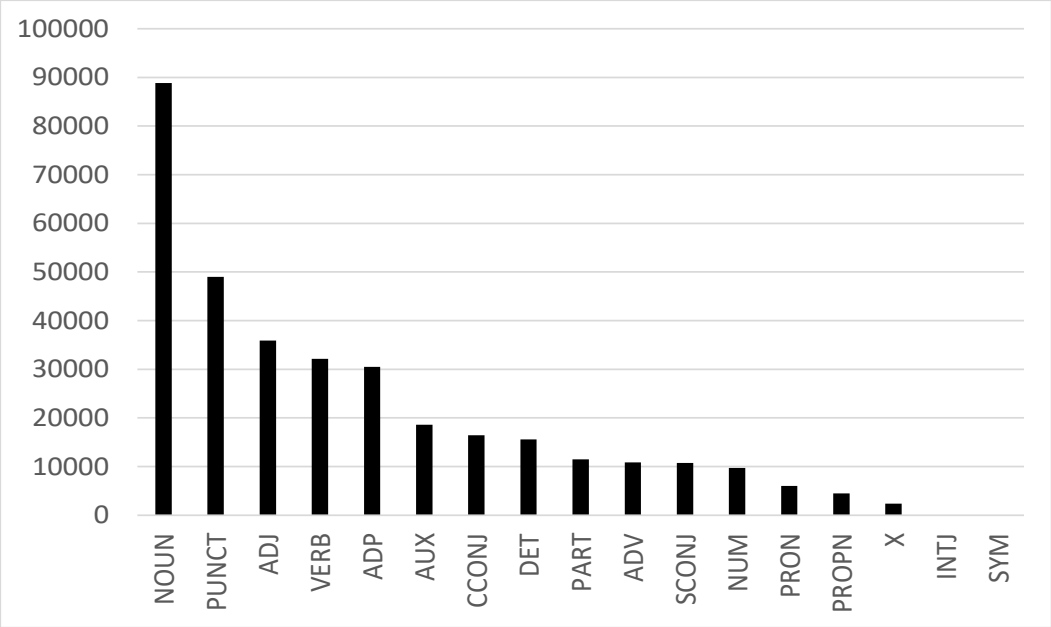
	Документ(и)	токена	речи	реченица	јединств. токена
1	Орвелова *1984.* (српски превод)	108.137	96.026	6677	18.050
2	Исечак из романа *Добри војник Швејк* Јарослава Хашека (српски превод)	4.122	3.347	266	1.475
3	Исечак из корпуса старих српских романа (1840-1920)	5.118	4.236	304	2.093
6	Новински чланци о поплавама у Србији 2014. године	4.672	3.813	288	1.741
7	Српски уџбеник из историје	6.596	5.287	331	2.622
8	Српко-енглески корпус права, финансија, образовања и здравља	214.159	193.643	9597	29.470
	Укупно	**342.804**	**306.352**	**17.463**	**34.334**

[1] https://live.european-language-grid.eu/catalogue/corpus/9295, приступљено 11. марта 2022.

[2] https://universaldependencies.org/u/pos, приступљено 11. марта 2022.

[3] http://www.korpus.matf.bg.ac.rs/SrpLemKor/tagset.html, приступљено 11. марта 2022.

За потребе експеримента, све реченице корпуса су насумично промешане, како за иницијално тестирање, тако и обучавања и евалуације касније, а за скуп ознака је који се посматра је узет *Universal POS*, чије су фреквенције ознака у корпусу приказане испод (Илустрација 12). Највећи број токена чине именице (*NOUN*), потом интерпункција (*PUNCT*), придеви (*ADJ*), глаголи (*VERB*) и предлози (*ADP*), те помоћни глаголи (*AUX*), напоредни и зависни везници (*CCONJ* и *SCONJ*), придевске и остале заменице (*DET* и *PRON*), речце (*PART*), прилози (*ADV*) и бројеви (*NUM*), док су најмање бројне властите именице (*PROPN*), узвици (*INTJ*), симболи (*SYM*) и остали, некатегорисани токени (*X*).



Илустрација 12: Расподела фреквенција ознака скупа *Universal POS* у *SrpKor4Tagging* анотираном корпусу.

5.3.2 Одабир тагера

Како би се максимално искористио утицај композитне архитектуре, потребна је темељна инспекција кандидата-тагера, а траже се они што бољи (да имају добре перформансе као самостални тагери) и што јединственији (јер нема смисла комбиновати резултате тагера који производе сличне резултате). Са друге стране, тагера треба да буде што више како би се постигли оптимални резултати у комбиновању (Aliwy, 2015).

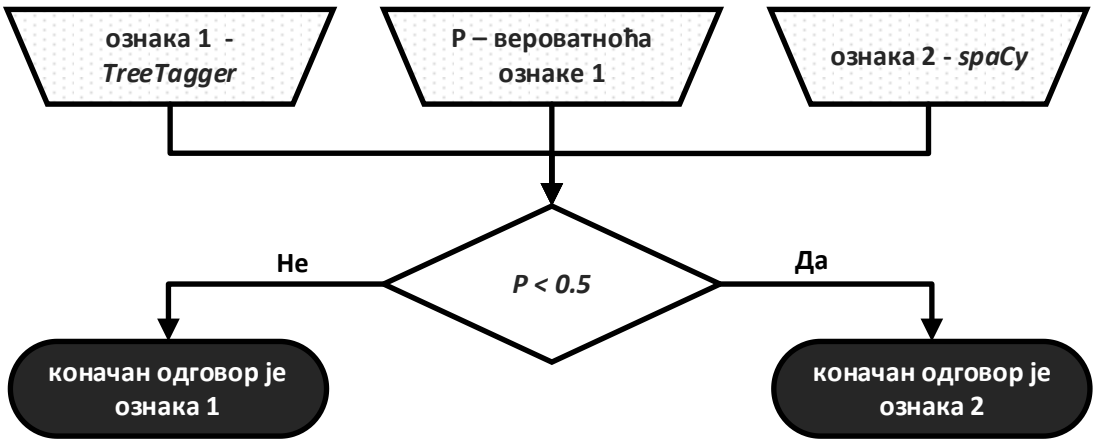
За српски језик, претходно поменути *TreeTagger* (одељак 5.1) се дуго сматрао оптималним приступом тагирања (Utvić, 2011), јер се директно ослања се на доступне и богате лексичке податке (Vitas & Krstev, 2012), али је деловало да је *spaCy*[4], који користи савременију технологију машинског учења, обећавајућа алтернатива (Šandrih, et al., 2019). Експерименти са другим тагерима, попут оног који је део

[4] https://spacy.io/, приступљено 31. марта 2022.

библиотеке *Natural Language Toolkit* (*NLTK*) (Bird, et al., 2009), нису произвели задовољавајуће резултате (Milovanović & Stanković, 2020). Станфордов нови систем *Stanza* (Qi, et al., 2020), који је наследник популарног софтвера *StanfordNLP* (Manning, et al., 2014), заснован је на условним насумичним пољима (*Conditional random fields*) и такође користи додатне, унапред обучене ресурсе попут модела за угњеждавање речи (*word embedding*) и стабала синтаксне зависности, како би се постигле побољшане перформансе. Ни он, међутим, није показао вансеријске резултате на задатку обележавања врсте речи за српски језик.

Неки од ових тагера, наиме *TreeTagger* и *spaCy* тагер за српски језик су поређени директно један против другог (Stanković, et al., 2020) и испоставило се да оба имају одређене предности и мане. Док *spaCy*-јев савремени приступ обуци заснован на рекурентним неуронским мрежама (у даљем тексту РНН) даје благу предност када постоји велико преклапање токена између скупова за обучавање и тестирање, он има слабији учинак у односу на *TreeTagger* када покушава да додели ознаку непознатим речима (речима које се не налазе у скупу за обучавање). Када се *TreeTagger* нађе у таквој ситуацији, врши се претрага речника, међутим, *spaCy* нема такву опцију у основном облику. Са тим на уму, намеће се идеја и да би комбиновањем ових тагера могло доћи до оптималних резултата, но ниједно од ових истраживања се није бавило тиме колико се одговори тагера међусобно разликују, што је, као што је већ поменуто, кључно за успешност паралелне композитне архитектуре (Aliwy, 2015).

Прелиминарна анализа поређења резултата ова два тагера састојала се од њиховог обучавања на 90% реченица корпуса *SrpKor4Tagging*, и тестирања на преосталих 10%. Том приликом је закључено да ови тагери не дају само различите, већ и јединствено тачне одговоре (добра ознака додељена од стране једног тагера, док је други није погодио). Наиме, *TreeTagger* постиже тачност од око 95.1%, при чему погађа 8% ознака које *spaCy* погрешно додељује, а, са друге стране, *spaCy* постиже тачност од око 94.5% и погађа 5% ознака које *TreeTagger* погрешно додељује. За детаљнију анализу, направљен је једноставан алгоритам, који као улаз има ознаку коју додељује *TreeTagger*, вероватноћу те ознаке (која је доступна као нуспојава тога што *TreeTagger* користи ХММ) и ознаку коју додељује *spaCy (*Илустрација 13*).*



Илустрација 13: Једноставан алгоритам који на основу ознака које додељују *TreeTagger* и *spaCy*, као и вероватноће ознаке коју је доделио *TreeTagger* додељује јединствену ознаку.

Уколико је вероватноћа коју је *TreeTagger* доделио својој ознаци већа или једнака 0.5, та ознака се узима као коначна, док се у супротном се узима ознака коју је доделио *spaCy*. Применом овог алгоритма над излазима тагера за 10% *SrpKor4Tagging* корпуса (ознаке скупа *Universal POS*) се добија нова, композитно-креирана листа ознака, која постиже тачност од око 95.6%, што надмашује резултате остварене од оба тагера засебно. Ипак, претходно је поменуто да је два тагера премало за композитну архитектуру, па је као следећи корак уследило проширивање листе тагера-кандидата.

За потребе тестирања које би одредило који тагери имају највећи потенцијал у композитној средини обучено је, поред поменута два (а на истом скупу података) још три додатна тагера: претходно поменути *Stanza* и *NLTK* тагери, као и *RNNTagger* (Schmid, 2019), унапређена верзија *TreeTagger*-а заснована на РНН. *Stanza*, која захтева предобучаване моделе за угњеждавање речи, опскрбљен је јавно доступним моделом за српски језик[5].

За одређивање који су тагери најбољи кандидати, осмишљен је тест заснован на мерењу међусобне ентропије – *Cross entropy loss* (*CEL*) одговора свих тагера, као и на мерењу ентропије између одговора сваког појединачног тагера и тачних ознака, при чему је ентропија између низова вероватноћа рачуната као:

$$\mathrm{CEL}(x, y) = -\sum_{i=1}^{n} x_i \log y_i$$

где је x један низ вероватноћа, y други низ вероватноћа, n укупан број њихових елемената, а i је индекс тих елемената. Тест је имао за циљ да детектује тагере који дају исувише сличне или исувише нетачне одговоре. Уз ентропију је, као додатна предострожност, мерена и могућност тагера да дају јединствено тачне одговоре (тачне одговоре које није погодио ниједан други тагер, у даљем тексту ЈТО).

С обзиром на то да је за потребе мерења ентропије пожељно да тагери одају и вероватноће својих избора, излази свих тагера су припремљени тако да та информација буде доступна током тагирања тест скупа. У овом одељку је већ поменуто да *TreeTagger* има ту могућност, а исто важи и за *NLTK* тагер, заснован на *мултиномном наивном бајесовском класификатору*. За потребе исписивања вероватноћа за *spaCy, Stanza* и *RNNTagger*-а, а с обзиром на то да су сва три заснована на РНН архитектури, било је потребно само процесирати њихове одговоре коришћењем софтмакс функције, како би се утврдиле тачне вероватноће. Оно што је тражено приликом овог тестирања су тагери који ће имати што већу *погодност*, израчунату као разлику просечне ентропије у односу на остале тагере и ентропије у односу на тачне ознаке, а уз то је било пожељно и да тагери имају што већи број ЈТО. Из резултата теста (Табела 3), видљиво је да *NLTK* и *RNNTagger* имају најмању стопу погодности (негативну), а уз то и јако мали број ЈТО: нула и седамнаест понаособ, те су они надаље елиминисани као тагери-кандидати.

[5] https://huggingface.co/stanfordnlp/stanza-sr/tree/main/models/pretrain, приступљено 31. марта 2022.

Табела 3: Поређење ентропије одговора припремљених тагера међусобно (горе) и просечна ентропија, ентропија у односу на тачне ознаке, разлика та два резултата и број јединствених тачних одговора, ЈТО (доле).

	TreeTagger	*spaCy*	*RNNTagger*	*Stanza*	*NLTK*
TreeTagger		6.15	6.81	6.40	7.79
spaCy	6.15		6.09	3.35	10.14
RNNTagger	6.81	6.09		5.27	5.82
Stanza	6.40	3.35	5.27		9.93
NLTK	7.79	10.14	5.82	9.93	
просек	6.79	6.43	6.00	6.24	8.42
тест скуп	4.85	5.19	6.84	5.71	9.05
погодност	1.94	1.24	**-0.84**	0.53	**-0.63**
ЈТО	1164	538	**17**	391	**0**

Како је листа кандидата спала на само три члана, тражени су алтернативни начини да се она даље прошири. Као решење је произишла двосмерна обука тагера, популаризована кроз феномен *BERT*-а. Идеја је била да се сви тагери обучавају не само на припремљеном корпусу, већ и на његовој инвертованој варијанти, како би се добили нови, потенцијално јединствени тагери. Ти *инвертовани* тагери (означени са [*И*]) би по обучавању исправно тагирали инвертован низ токена, а произведена ниска ознака би поново била инвертована како би одговарала ординацији оригиналних токена (Stanković, et al., 2022). Но, како би се утврдила њихова погодност за композитну архитектуру тестирани су једнако као и самостални тагери, а резултати овог теста су приказани у наставку (Табела 4).

Табела 4: Поређење ентропије одговора припремљених тагера међусобно (горе) и просечна ентропија, ентропија у односу на тачне ознаке, разлика та два резултата и број ЈТО и двосмерног ЈТО (доле).

	TreeTagger	*spaCy*	*Stanza*	*TreeTagger* [*И*]	*spaCy* [*И*]	*Stanza* [*И*]
TreeTagger		6.15	6.40	1.54	6.18	6.27
spaCy	6.15		3.35	6.41	4.12	4.00
Stanza	6.40	3.35		6.53	4.35	3.19
TreeTagger [*И*]	1.54	6.41	6.53		6.43	6.18
spaCy [*И*]	6.18	4.12	4.35	6.43		5.06
Stanza [*И*]	6.27	4.00	3.19	6.18	5.06	
просек	5.31	4.81	4.76	5.42	5.23	4.94
тест скуп	4.25	5.19	5.71	4.58	5.17	6.05
погодност	1.06	-0.38	-0.95	0.84	0.06	-1.31
ЈТО	57	59	57	42	59	48
ДЈТО	444	1038	1002	394	1064	695

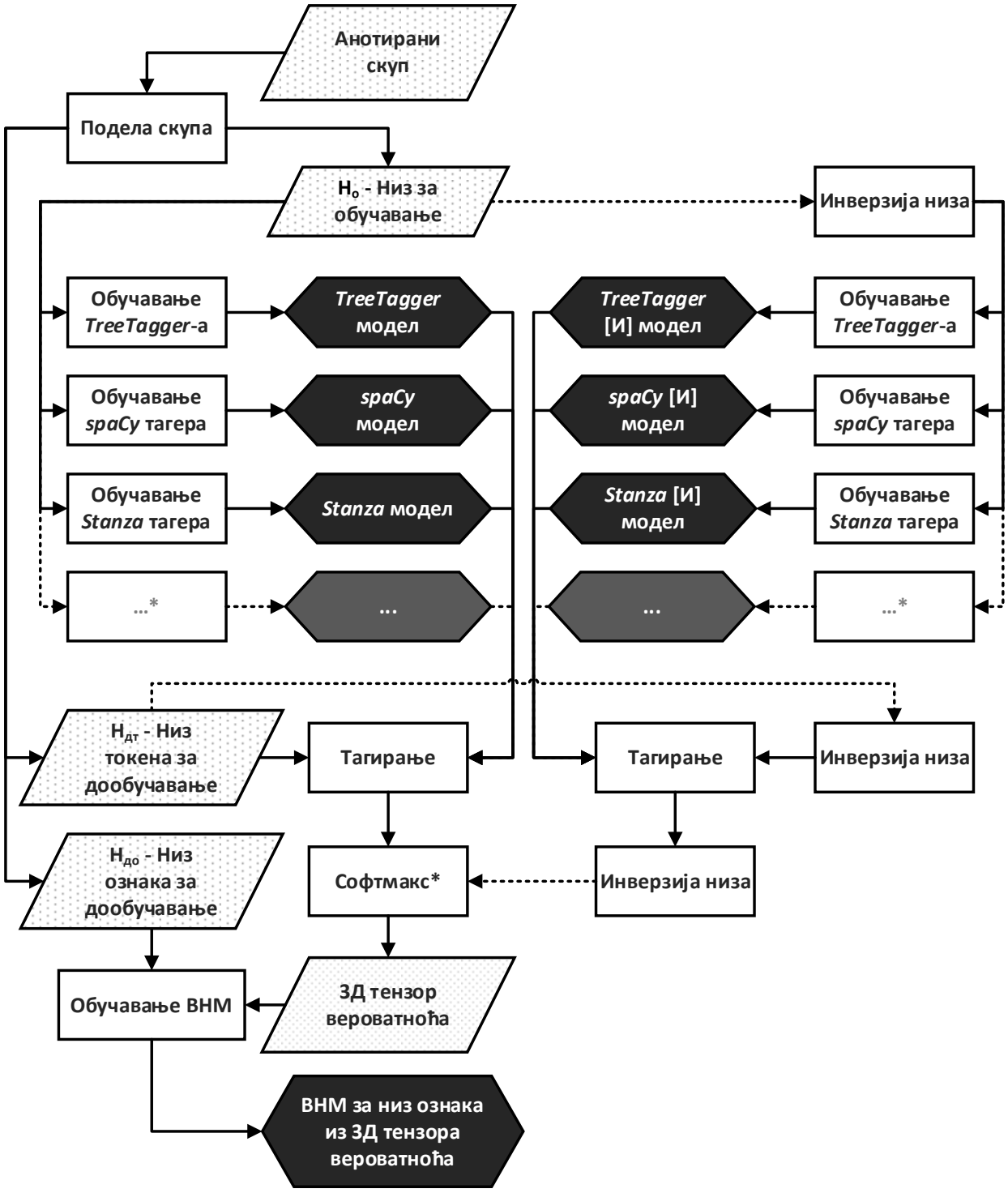
Из ових резултата је видљиво да је стандардна девијација ентропија (и степена погодности) много мања у односу на претходни тест (Табела 3, Табела 4), а сви кандидати имају поприличан број ЈТО и ДЈТО (двосмерни ЈТО тј. ЈТО у односу на инвертованог парњака и обратно) бодова, па су стога сматрани подједнако валидним да као самостални буду укључени у композитну структуру.

5.3.3 Алгоритми за обучавање и тагирање

Композитно обучавање, тагирање и евалуација, а поготово оно које је двосмерно-проширено свакако захтева посебне алгоритме, па су тако они развијени и за потребе овог експеримента.

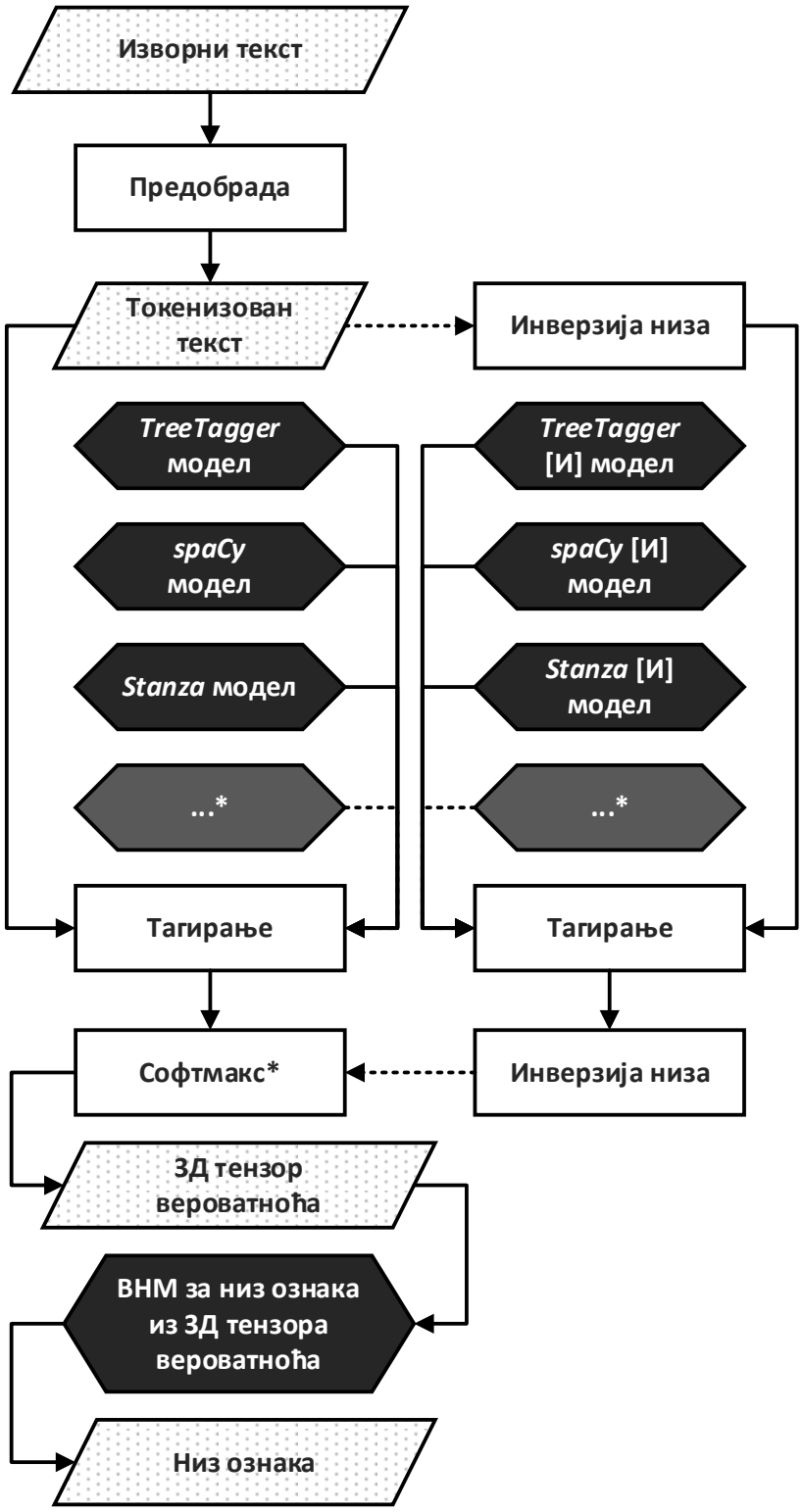
Први алгоритам се састоји од двосмерног обучавања изабраних тагера и креирање скупа за дообучавање наслаганог класификатора на основу вероватноћа ознака које самостални тагери производе над скупом за дообучавање. Можемо да га поделимо на четири корака (Илустрација 14):

1. Формирање скупова за обучавање и дообучавање – Анотирани скуп реченица које планирамо да користимо за обучавање композитног тагера најпре промешамо и поделимо на два дела: већи—низ за обучавање ($Н_о$) и мањи—низ за дообучавање, који даље вертикално делимо на низ токена ($Н_{дт}$) и низ ознака ($Н_{до}$). Резултат првог корака су ова три низа;
2. Двосмерно обучавање самосталних тагера – Низ који се користи за обучавање, $Н_о$ треба ископирати и копију инвертовати, како би се обучили и стандардни и инвертовани тагери (обучавање се врши на уобичајен начин). Такође, треба напоменути да је овај приступ скалабилан, те да се може обучити било који број тагера, докле год они имају опцију да доделе ознакама вероватноће. Резултат другог корака су модели тагера обучени на већем делу анотираног корпуса и њихови инвертовано-обучени парњаци;
3. Креирање тродимензионалног тензора вероватноћа – Овај тензор се добија конкатенацијом вероватноћа свих ознака за све тагере, а вероватноће се добијају тагирањем другог низа креираног у првом кораку, $Н_{дт}$. Дакле, сваки тагер (са тим да инвертовани тагери тагирају инвертован низ, па се резултати поново инвертују у оригиналан редослед) обрађује низ токена, а резултоване вероватноће (за сваки тагер, сваку ознаку и сваки токен) се конкатенирају у један тензор који је резултат овог корака;
4. Обучавање наслаганог класификатора – Тензор добијен у претходном кораку се користи као скуп за обучавање додатног класификатора заједно са трећим низом из првог корака, $Н_{до}$ (који у овом случају представљају класе које класификатор треба да научи да предвиђа). Класификатор који се обучава заснива се на ВНМ, а чини га један *перцептрон* (два свеповезана слоја неурона). Резултат овог, последњег корака је, дакле, класификатор који на основу добијених вероватноћа за сваку ознаку од стране сваког тагера предвиђа једну, коначну ознаку.



Илустрација 14: Архитектура за обучавање проширивог двосмерног композитног тагера, где десна половина представља двосмерно проширење, а могућа проширења у виду додатних тагера су обележена са …*.

У складу са описаном архитектуром за обучавање креирана је и архитектура за тагирање, која на основу обучених самосталних модела тагера и обученог наслаганог класификатора (који су резултат комплетног обучавања) обезбеђује одговарајући низ ознака за неки задати низ токена (Илустрација 15).



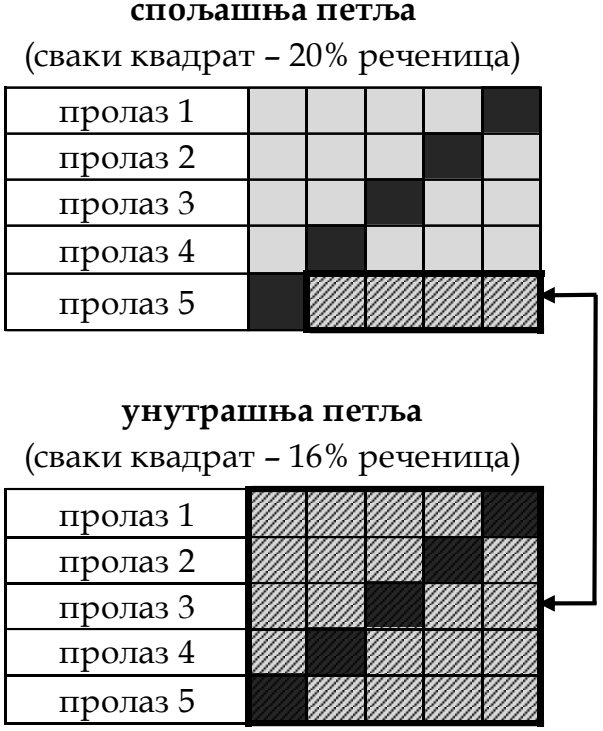
Илустрација 15: Архитектура за тагирање проширивог двосмерног композитног тагера, где десна половина представља двосмерно проширење, а могућа проширења у виду додатних тагера су обележена са …*.

Тагирање се одвија у три једноставна корака:

1. Предобрада – Први корак се састоји од токенизације изборног, сировог текста који желимо да тагирамо;
2. Генерисање карактеристика – Током другог корака сваки од обучених самосталних тагера сваком од задатих токена додељује вероватноће за сваку могућу ознаку. Те вероватноће се конкатенирају у јединствен тензор карактеристика;
3. Употреба наслаганог класификатора – Обучена ВНМ узима тензор вероватноћа из претходног корака и користи га да образује резултовани низ ознака који одговара низу токена добијеном у првом кораку.

5.3.4 Евалуација одређивања врсте речи

За потребе тестирања развијених алгоритама осмишљена је евалуација заснована на угњежденој петострукој унакрсној валидацији, од укупно 25 (пет пута пет) пролаза (Илустрација 16).



Илустрација 16: Визуелизација спољашње и унутрашње петље угњеждене петоструке унакрсне валидације.

За коначно обучавање и евалуацију тагера поново је коришћен *SrpKor4Tagging* корпус и скуп ознака *Universal POS*. У спољашњој петљи је за сваки од пет пролаза корпус подељен на скупове реченица у размери од четири према један тј. у сваком од пет пролаза 80% реченица је коришћено за обучавање модела који су тестирани на преосталих 20% реченица. Како и наслагани класификатор захтева одређени број реченица за дообучавање, та подела се обављала у унутрашњој петљи, где су 80% реченица предвиђених за обучавање поново груписане у размери четири према један

тако да се на већем делу се обучавају самостални тагери, а на мањем наслагани класификатор. Укратко, у сваком од 25 пролаза, шест самосталних тагера (три стандардна и три инвертована) се обучава на порцији од 64% реченица корпуса, на 16% се обучава наслагани класификатор, а на 20% се евалуирају резултати. За сваки од ових 25 уникатних пролаза, тестирана су и три претходно поменута начина композиције: гласање, бодовање и лицитирање.

Резултати су израчунати у виду Ф-мере, а рачунати су за сваки самостални тагер, као и четири композитна метода. Даље су приказани упросечени прикупљени резултати (Табела 5), а вреди напоменути да није било великих одступања међу њима. Такође, приказани су и минимални и максимални резултат од свих пролаза за сваки тагер, те количник та два броја који представља *стабилност* тагера у односу на исечак корпуса на којој је обучаван. У табели је приказан и упросечен резултат свих самосталних тагера, као и најбољи резултат међу њима, који је узет као основица за евалуацију унапређења које доноси композитна архитектура.

Табела 5: Резултати евалуације свих самосталних тагера (горе), упросечени резултати и основица за мерење побољшања (најбољи резултати међу самосталним тагерима) (средина) и резултати које су постигли композитни тагери (доле).

	F-мера	**минимум**	**максимум**	**стабилност**
spaCy	0.9455	0.9436	0.9470	**0.9964**
spaCy[И]	0.9461	0.9428	0.9483	0.9942
TreeTagger	**0.9542**	0.9519	**0.9556**	0.9961
TreeTagger[И]	0.9534	**0.9522**	0.9540	0.9981
Stanza	0.9019	0.8741	0.9389	0.9310
Stanza[И]	0.8942	0.8431	0.9275	0.9090
просек	0.9325	0.9193	0.9445	0.9734
основица	0.9542	0.9522	0.9556	0.9964
гласање	0.9592	0.9536	0.9651	0.9881
лицитирање	0.9721	0.9697	0.9741	0.9955
бодовање	0.9747	0.9736	0.9756	0.9979
слагање	**0.9755**	**0.9751**	**0.9762**	**0.9989**

Из приказаних резултата се може видети између осталог да:

1. *TreeTagger* показује најбоље перформансе, у складу са претходним истраживањима за српски језик, укључујући најбољу просечну Ф-меру и највеће минималне и максималне резултате;
2. Инвертовани тагери показују резултате на нивоу стандардних тагера. *TreeTagger* надмашује *TreeTagger*[И] у просечној Ф-мери и највишој максималној вредности, али *TreeTagger*[И] има најбољу минималну вредност. *spaCy*[И]

показује већи просечан резултат за Ф-мере од *spaCy* тагера. Заједно са претходно утврђеном разликом у ентропији (Табела 4), ово додаје на значају инвертованим тагерима као самосталним и утврђује методу инверзије скупа за обучавање тагера као занимљиву, бар за српски језик.

Будући да је било пожељно тестирати како двосмерно обучавање утиче на побољшање резултата, за сваки пролаз су израчунати и резултати композитних метода једносмерно, дакле, игноришући инвертоване тагере при комбиновању. Ови резултати су приказани испод (Табела 6).

Табела 6: Упросечени резултати самосталних тагера и основица за мерење побољшања истоветно као у Табела 5 (горе), као и резултати које су постигли композитни тагери игноришући инвертоване тагере (доле).

	F-мера	**минимум**	**максимум**	**стабилност**
просек	0.9325	0.9193	0.9445	0.9734
основица	0.9542	0.9522	0.9556	0.9964
гласање	0.9537	0.9490	0.9612	0.9873
лицитирање	0.9600	0.9589	0.9611	0.9977
бодовање	0.9691	0.9668	0.9718	0.9949
слагање	**0.9714**	**0.9700**	**0.9721**	**0.9978**

Како би се лакше поредили, приказани резултати (Табела 5 и Табела 6) су даље обрађени у циљу директног поређења између два приступа (Табела 7), а поређен је степен смањења грешке, израчунат као:

$$\text{Степен смањења грешке} = \frac{a_1 - a_0}{1 - a_0}$$

где је a_0 основица, F-мера коју је остварио најбољи самостални тагер, а a_1 F-мера коју је остварио неки композитни приступ.

Табела 7: Степен смањења грешке који се јавља као последица коришћења инвертованих тагера за све претходно описане мере.

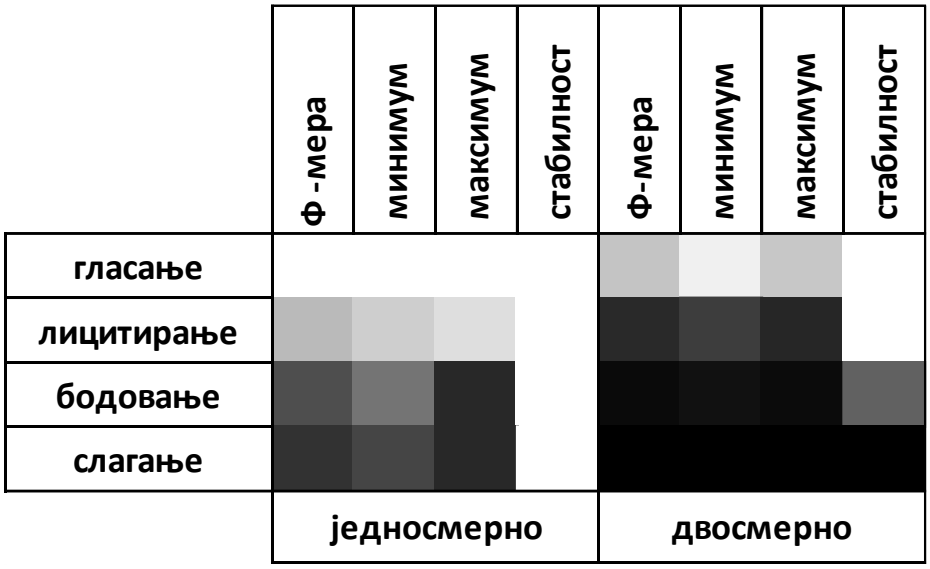
	F-мера	**минимум**	**максимум**	**стабилност**
гласање	11.65%	7.21%	19.32%	0.08%
лицитирање	**30.17%**	**26.28%**	**33.42%**	-0.22%
бодовање	17.72%	10.53%	26.20%	**0.31%**
слагање	14.35%	10.75%	17.33%	0.10%

Из принуђеног се види значајна предност у коришћењу инвертованих тагера у композитној средини (или просто удвостручавања броја тагера који генеришу карактеристике). Степен смањења грешке за упросечену Ф-меру износи од 11.65% до

чак 30.17%, у зависности од приступа, а са њим се повећавају минимум, максимум и (у највећем броју случајева) стабилност. Поређење резултата које су остварили композитни тагери уопште у односу на стандардне, такође изведено из претходно приказаних резултата (Табела 5 и Табела 6) приказано је табеларно у наставку (Табела 8) и додатно визуализовано, у виду топлотне мапе (Илустрација 17).

Табела 8: Степен смањења грешке који пружају једносмерно (горе) и двосмерно засновани композитни тагери (доле) за све претходно описане мере.

	F-мера	**минимум**	**максимум**	**стабилност**	**двосмерно**
гласање	-1.34%	-13.33%	14.35%	-256.73%	не
лицитирање	12.67%	9.23%	18.50%	35.66%	не
бодовање	32.44%	26.22%	41.37%	-44.61%	не
слагање	**37.47%**	**35.06%**	**41.37%**	**39.28%**	не
гласање	10.87%	2.93%	21.40%	-234.91%	да
лицитирање	39.04%	36.61%	41.67%	-26.95%	да
бодовање	44.72%	44.77%	45.05%	42.38%	да
слагање	**46.51%**	**47.91%**	**46.40%**	**68.33%**	да



Илустрација 17: Визуелизација смањења степена грешке који пружају композитни тагери за све претходно описане мере, где тамнија боја представља веће смањење грешке (бољи резултат), а бела боја означава да нема смањења грешке (нема побољшања).

Из овога се може видети неколико ствари:

1. Готово сви композитно-засновани тагери надмашују најбољи самостални тагер (основицу) у свим категоријама, а једини изузеци су једносмерно-засновани метод гласања, што се тиче упросечене и минималне постигнуте Ф-мере. Сви остали методи надмашују основицу

и показују степен смањења грешке и до 46% (коришћењем методе слагања), што недвосмислено даје потврдан одговор на ИП1:

Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и њиховим излазним вероватноћама адекватна метода у области обраде природних језика?

2. Потврђује се да претходно поменута констатација да двосмерно-засновани тагери показују боље резултате од једносмерних, што је сада посебно уочљиво на приказаној илустрацији (Илустрација 17) и даје потврдан одговор на ИП2:

Да ли се композитни интелигентни системи засновани на паралелним моделима могу надограђивати – и да ли већи број активних паралелних модела побољшава квалитет целокупног система?

3. Метод слагања показује најбоље резултате од свих композитно-заснованих тагера што се види на резултатима Ф-мере, минимума, максимума и (посебно) стабилности. Слагање показује најбоље резултате у свим категоријама и међу једносмерно и двосмерно-заснованим композитним тагерима.

Закључено је, дакле, да је метод паралелног приступа обраде текста погодан за српски језик, и да се коришћењем композитних модела постиже велико побољшање у односу на самосталне тагере. Пошто је такав систем скалабилан, нови модели се лако могу додати што ову методу чини погодном за каснију надоградњу и унапређење. Додатно, метод слагања, кроз обучавање наслаганом класификатора над излазима самосталних модела се истиче као оптималан приступ приликом оваквог, паралелног, компоновања.

6
Паралелни језички модели у моделовању мини-језика

6.1 Удаљено читање, анализа ауторства и моделовање стила као мини-језика

Удаљено читање је метода квантитативне анализе текста у студијама књижевности и подразумева коришћење рачунарства у обради великих збирки књижевних текстова, све са циљем да се допуне класичне методе изучавања (Moretti, 2000). Ова парадигма је нашла корист пре свега у истраживању великих корпуса *на даљину* тј. сагледавању неких специфичних ширих карактеристика унутар текстова, што омогућава објективније откривање претходно незапажених информација и образаца у њима. Методолошка новина, која се састоји од коришћења узорака текста, статистике и паратекста метаподатака, омогућује истраживачима да сазнају више о текстовима чак и без њиховог детаљног читања.

Ова метода је, између осталог, имала утицај на реорганизацију *анализе ауторства*, задатка ОПЈ који се бави екстракцијом информација о аутору из текста. Док су се рани методи анализе ауторства заснивали на лингвистици и аритметици, на пример, још у деветнаестом веку су рачунате фреквенције речи у неким потенцијалним делима Шекспира, како би се утврдило да ли је он уистину њихов аутор (Mendenhall, 1887), са увођењем рачунарских метода (а поготову машинског учења) процес се умногоме аутоматизује и олакшава, а постигнути резултати постају све бољи и бољи (Stamatatos, 2009).

Иако се савремена истраживања међусобно методолошки и технички разликују, сва се ослањају на запажању да су речи употребљене у тексту добри показатељи *стила* које аутор користи (El Manar El Bouanan & Kassou, 2014), а то се посебно односи на оне најфреквентније употребљаване, у складу са *Зипфовим закон*ом. У том случају, *стилске*

удаљености између текстова се могу рачунати, на пример, као еуклидске раздаљине вектора фреквенција речи два текста, мада примена *Баровсове Делте* (*Burrows Delta*) (Burrows, 2002) и њених деривација попут *косинусне делта удаљености* (Evert, et al., 2015) даје видно боље резултате у пракси. Оваква мерења стила и стилске удаљености, неопходна при решавању задатка одређивања и провере ауторства, заснована су у *стилометрији*, методи статистичке анализе текстова и квантификације стила, специфичног *мини-језика* на којим су ти текстови писани.

Моделовање стила, као најједноставнији специфичан подзадатак моделовања језика (слично тагирању врстом речи у ОПЈ) је веома погодан терен за тестирање нових приступа и техника. Управо због тога, методе *паралелног моделирања језика*, неопходне за развој композитне граматике (или псеудограматике) српског језика, тестиране су и развијане на овом задатку (Škorić, et al., 2022).

6.2 *Стилометрија и одређивање ауторства*

Принцип на коме се, дакле, заснива стилометрија је да сваки појединачни аутор има одређени стил писања којим се служи, што у теорији омогућава разликовање докумената различитих аутора и решавања проблема потврде и одређивања ауторства.

Перформансе појединих стилометријских метода уско зависе од избора језичких карактеристика као релевантних маркера стила, а одређивање које карактеристике је најбоље користити (за решавање одређених задатака) и како их треба екстраховати је и даље предмет дебате. Најранији приступи су се ослањали искључиво на речи и испитивали су разлике у њиховој употреби код појединих аутора (Mendenhall, 1887) или на листе најфреквентнијих речи (Burrows, 2002), док су потоња експериментисала са алтернативним карактеристикама попут лема и врста речи (Rybicki & Eder, 2011; Eder & Górski, 2022). Осим карактеристика, упитни су и методи израчунавања удаљености између њих, те стратегије нормализације и различитост језичких породица (Eder & Górski, 2022).

Међу приступима који се не заснивају директно на фреквенцијама речи, данас се најчешће користе н-грами токена и граматичке категорије, али и неки комбиновани приступи, који су посебно занимљиви за тему овог рада. У паралелним архитектурама за одређивање ауторства могућа је (као и код тагирања врстом речи) употреба наслаганог, хетерогеног класификатора који комбинује независне класификаторе засноване на различитим приступима, а овакав приступ дакако надмашује оне добијене коришћењем самосталног класификатора (Stamatatos, et al., 2014; Akimushkin, et al., 2018; Weerasinghe & Greenstadt, 2020), једнако као и на задатку тагирања врстом речи. У раду (Segarra, et al., 2015) аутори такође показују да комплексне мреже суседних речи и фреквенције речи обухватају различите стилометријске аспекте те да њихова комбинација може преполовити стопу грешке постојећих метода.

Осим наведених, плитких репрезентација докумената, заснованих на скупу (врећи) речи (*bag of words*), недавне студије у одређивању ауторства истражују и контекстно-зависне репрезентације или карактеристике, обично екстраховане помоћу ВНМ (Kocher & Savoy, 2018; Salami & Momtazi, 2021).

6.3 Векторизација и угњеждавање докумената

Одређивање ауторства (као и провера ауторства) неизбежно подразумева поређење парова докумената различитих аутора, како би се утврдио степен сличности међу њима. Већ смо поменули да је најпопуларнији начин поређења докумената израчунавање удаљености између њихових векторских репрезентација које се могу добити на различите начине, о чему је такође било речи у одељку о стилометрији (6.2). Из тог разлога, када се дође до тога да су израчунате међусобне удаљености/сличности неке групе докумената, погодан начин за њихово представљање је симетрична, шупља матрица димензија *k*k* (где је *k* број докумената које анализирамо), која се обично назива *матрица угњежђења* или *матрица удаљености докумената* (Teofili, 2019). Када је матрица удаљености докумената припремљена, може бити од велике користи у анализи ауторства, под претпоставком да су документи између којих је најмања удаљеност у матрици писани од стране истог аутора, наравно, у случају да су представљене удаљености засноване у стилометрији. Матрицу удаљености представљамо као:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \cdots & a_{k,k} \end{bmatrix} = \begin{bmatrix} 0 & * & \cdots & * \\ * & 0 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 0 \end{bmatrix}$$

где је $*$ нека арбитрарна нумеричка удаљеност.

6.3.1 Векторизација докумената заснована на фреквенцијама (н-грама) токена

Уколико желимо да креирамо поменуту матрицу удаљености засновано на најуобичајенијој методи, фреквенцијама речи у документу, можемо користити *Stylo* пакет за *R* програмски језик, развијен кроз сарадњу истраживача Института за пољски језик и Јагиелноског универзитета у Кракову, и Универзитета у Антверпену (Eder, et al., 2016), а са циљем да се олакша и аутоматизује стилометријска анализа текстова.

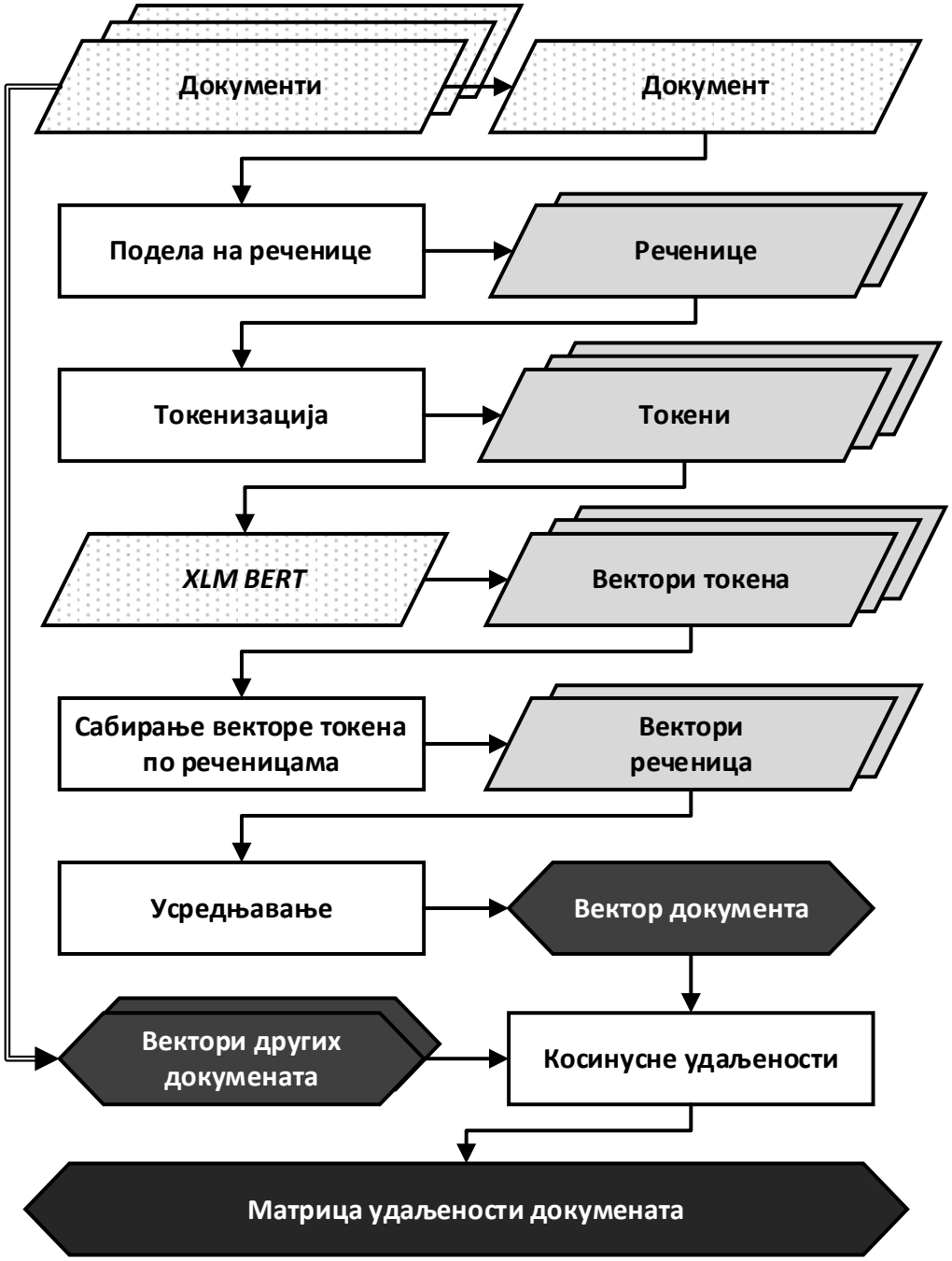
Stylo подржава и н-грамску трансформацију докумената, па тако можемо помоћу њега добити не само листе најфреквентнијих речи (токена), него и листе најфреквентнијих н-грама. Удаљености између докумената се могу рачунати помоћу различитих предефинисаних мера удаљености, при чему аутори препоручују претходно поменуту косинусну делта удаљеност (Evert, et al., 2015), познату и под називом *Вурцбушка удаљеност*. Документи се, даље, могу аутоматски скратити тако да буду истих дужина или поделити на сегменте једнаке дужине, што је оптимално за правилно одређивање удаљености. Листе најфреквентнијих речи се, такође, могу скратити на неку одабрану величину, при чему се за анализу ауторства препоручује број између 100 и 1000 у зависности од употребе и укупног броја кандидата у листи.

Дакле, помоћу овог пакета можемо неки документ (или групу докумената) поделити на сегменте једнаке дужине (оптимално за анализу ауторства), извршити њихову н-грамску трансформацију (опционо), израчунати листу најфреквентнијих н-грама (или униграма), па потом на основу тих листа израчунати удаљености између њих и креирати матрицу удаљености (Илустрација 18).

Илустрација 18: Креирање матрице удаљености докумената засновано на векторима фреквенције токена и косинусној делта удаљености између њих. Испрекидана линија представља опционо прескакање н-грамске трансформације, а двострука линија представља генерисање вектора за сва документа.

6.3.2 Векторизација докумената заснована на *BERT* моделима

Као што смо већ споменули векторизација докумената за потребе стилометријске анализе не мора се нужно ослањати на листе најфреквентнијих речи. У раду (Iyer & Vosoughi, 2020) аутори експериментишу са коришћењем *BERT* модела при креирању матрице удаљености докумената за потребе анализе ауторства, и развијају сопствени алгоритам заснован на претходно поменутом мултилингвалном, *XLM BERT*-у (одељак 2.4.2) и косинусној удаљености између вектора (Илустрација 19).



Илустрација 19: Креирање матрице удаљености докумената засновано на векторима угњежђења додељених од стране *BERT* модела и косинусне удаљености између њих. Двострука линија представља генерисање вектора свих докумената.

Према овој схеми, документи који су предмет анализе се најпре поделе на реченице, за шта се користи неколико грубо припремљених регуларних израза. Добијене реченице се потом токенизују токенизатором развијеним специјално за *BERT*, а који су развили истраживачи из Гугла, при чему се све реченице дуже од 512 токена скраћују на ту дужину, како би их модел могао обрадити. *XLM BERT* додељује сваком токену сваке реченице вектор величине 768, а додељени вектори се потом сабирају унутар реченица како би се произвели реченични вектори исте величине. Сви реченични вектори за сваки документ се усредњавају (сабирају се и резултовани

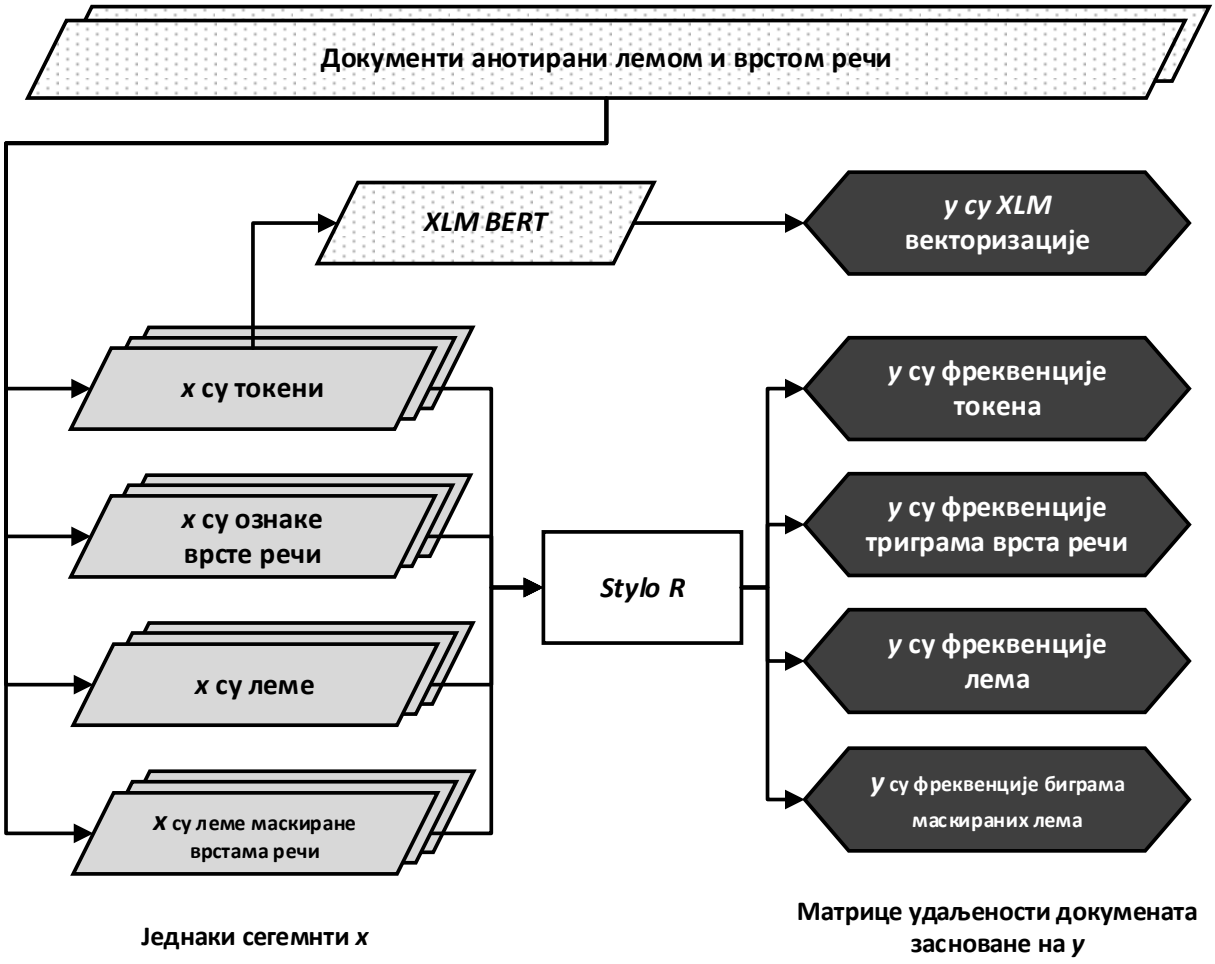
вектор се дели са укупним бројем реченица) тако да сваки документ добија сопствени вектор величине 768. Коначно, између k вектора документа, $\overrightarrow{v_i}, \overrightarrow{v_j}$, $i, j \in \overline{\{1, k\}}$, израчунавају су косинусне удаљености:

$$d_{i,j} = \frac{\|\overrightarrow{v_i}\| \cdot \|\overrightarrow{v_j}\|}{\langle \overrightarrow{v_i}, \overrightarrow{v_j} \rangle}$$

а добијене вредности се смештају у матрицу удаљености докумената.

6.3.3 Паралелне матрице удаљености докумената

Уколико желимо да комбинујемо резултате више различитих приступа угњеждавања докумената, то можемо чинити и на нивоу самих матрица. Узмимо за пример пет јединствених матрица удаљености направљених коришћењем пет различитих метода, конкретно, једну користећи претходно поменути приступ заснован на *XLM BERT*-у (Iyer & Vosoughi, 2020) и четири заснована на *Stylo R* пакету и различитим репрезентацијама докумената (оригинални документ, лематизован документ, документ у којем су речи замењене ознакама врстама речи, и лематизован документ у којем су речи селективно замењене), као што је приказано испод (Илустрација 20).



Илустрација 20: Креирање паралелних матрица удаљености докумената.

На основу n квадратних матрица удаљености k различитих докумената, чије вредности припадају скупу реалних бројева ($D_1, D_2, \dots, D_n \in M_k(\mathbb{R})$), и које се могу записати као:

$$\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & a_{1,k}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & \cdots & a_{2,k}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1}^{(1)} & a_{k,2}^{(1)} & \cdots & a_{k,k}^{(1)} \end{bmatrix}, \begin{bmatrix} a_{1,1}^{(2)} & a_{1,2}^{(2)} & \cdots & a_{1,k}^{(2)} \\ a_{2,1}^{(2)} & a_{2,2}^{(2)} & \cdots & a_{2,k}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1}^{(2)} & a_{k,2}^{(2)} & \cdots & a_{k,k}^{(2)} \end{bmatrix}, \dots, \begin{bmatrix} a_{1,1}^{(n)} & a_{1,2}^{(n)} & \cdots & a_{1,k}^{(n)} \\ a_{2,1}^{(n)} & a_{2,2}^{(n)} & \cdots & a_{2,k}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1}^{(n)} & a_{k,2}^{(n)} & \cdots & a_{k,k}^{(n)} \end{bmatrix}$$

дефинишемо композитну матрицу D_m

$$D_m = D_m(D_1, D_2, \dots, D_n) = \begin{bmatrix} b_{1,1}^m & b_{1,2}^m & \cdots & b_{1,k}^m \\ b_{2,1}^m & b_{2,2}^m & \cdots & b_{2,k}^m \\ \vdots & \vdots & \ddots & \vdots \\ b_{k,1}^m & b_{k,2}^m & \cdots & b_{k,k}^m \end{bmatrix}$$

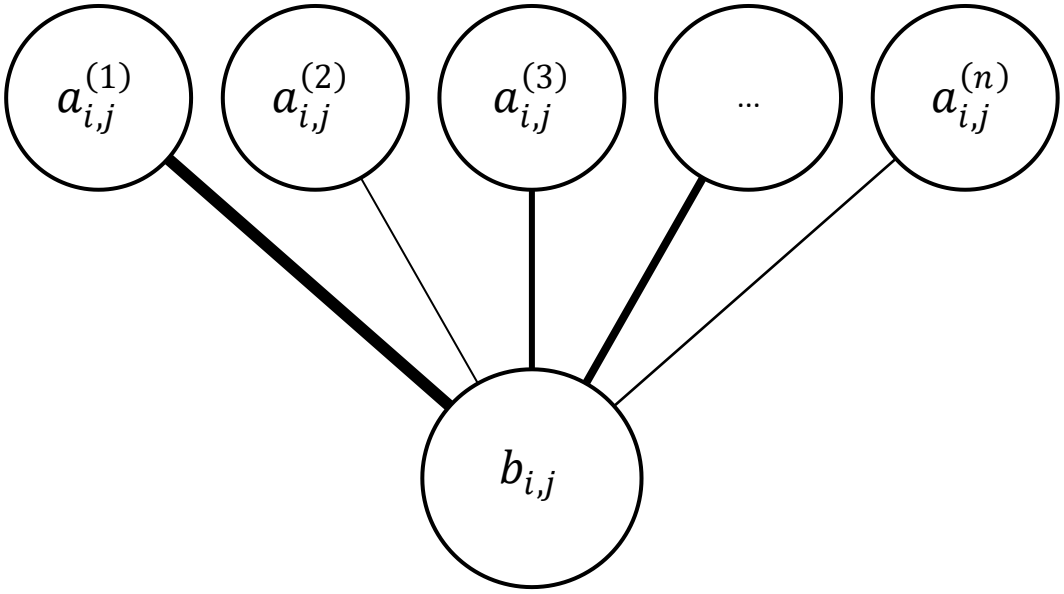
где сваки њен елемент $b_{i,j}^m, i, j \in \overline{\{1,k\}}$ може бити генерисан помоћу различитог метода m, на пример (Škorić, et al., 2022):

$$b_{i,j}^m = \begin{cases} \dfrac{a_{i,j}^{(1)} + a_{i,j}^{(2)} + \cdots + a_{i,j}^{(n)}}{n}, i, j \in \overline{\{1,k\}}, & m \text{ је аритметичка средина} \\ a_{i,j}^{(1)} * a_{i,j}^{(2)} * \dots * a_{i,j}^{(n)}, i, j \in \overline{\{1,k\}}, & m \text{ је производ} \\ \min\left(a_{i,j}^{(1)}, a_{i,j}^{(2)}, \dots, a_{i,j}^{(n)}\right), i, j \in \overline{\{1,k\}}, & m \text{ је минимум} \\ \max\left(a_{i,j}^{(1)}, a_{i,j}^{(2)}, \dots, a_{i,j}^{(n)}\right), i, j \in \overline{\{1,k\}}, & m \text{ је максимум} \\ \sqrt{\left(a_{i,j}^{(1)}\right)^2 + \left(a_{i,j}^{(2)}\right)^2 + \cdots + \left(a_{i,j}^{(n)}\right)^2}, i, j \in \overline{\{1,k\}}, & m \text{ је векторска норма} \end{cases}$$

Пример 6: Неки од могућих метода за комбиновање матрица удаљености докумената и формуле за њихово израчунавање.

На овај начин добијамо пет нових, јединствених матрица удаљености које су истог облика као и оне од којих су настале, па се могу лако међусобно поредити како би се одредило који метод угњеждавања је најоптималнији.

Алтернативно, можемо користити и наслагани класификатор који узима матрице удаљености ($D_1, D_2, \dots, D_n \in M_k(\mathbb{R})$) као улаз и генерише нову матрицу истих димензија као излаз. За ту потребу би се низ ћелија на истом положају у основним матрицама $a_{i,j}^{(1)}, a_{i,j}^{(2)}, \dots, a_{i,j}^{(n)}$, користио као улаз за ВНМ, где би се као излаз производила вредност $b_{i,j}$, која би се налазила на одговарајућем месту у новој композитно-заснованој матрици (Илустрација 21).



Илустрација 21: Схема једноставне ВНМ, свеповезаног перцептрона са једним излазом који од n вредности удаљености производу нову јединствену вредност.

Оваква ВНМ би обучавањем дошла до одговарајућих тежина коју треба да носи свака од матрица удаљености (како би се произвели оптимални резултати), па би се тако она дефинисала и као

$$D_w = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,k} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k,1} & b_{k,2} & \cdots & b_{k,k} \end{bmatrix}$$

где је

$$b_{i,j} = \frac{a_{i,j}^{(1)} w^{(1)} + a_{i,j}^{(2)} w^{(2)} + \cdots + a_{i,j}^{(n)} w^{(n)}}{w^{(1)} + w^{(2)} + \cdots + w^{(n)}}, i, j \in \overline{\{1, k\}}$$

и где су $w^{(1)}, w^{(2)}, \ldots, w^{(n)} \in \mathbb{R}$ реалне вредности које одговарају тежинама везе између одговарајућих улазних чворова ВНМ и излазног чвора, као што је претходно приказано (Илустрација 21).

6.4 *Угњеждавање старих српских романа*

Методе описане у претходном одељку су тестиране на задатку одређивања ауторства старих (српских) романа у циљу одговора, између осталог, на следећа питања:

1. Који је метод репрезентације докумената на српском језику оптималан за задатак одређивања ауторства;
2. Да ли се на нивоу матрица могу комбиновати удаљености како би се побољшали резултати на задатку одређивања ауторства;

3. Да ли је наслагани класификатор оптималан метод комбиновања матрица удаљености ради унапређења резултата на задатку одређивања ауторства;
4. Да ли укључивање матрице удаљености засноване на *XLM BERT* моделу позитивно утиче на резултате приликом решавања поменутог задатка.

Ово је постигнуто реализацијом следећих корака:

1. Израдом матрица удаљености за пет различитих репрезентација документа и њихово тестирање једне против других на задатку одређивања ауторства. Репрезентације укључују: оригинални документ, лематизован документ, документ у коме су сви токени замењени њиховим ознакама врста речи из скупа *Universal POS*, и коначно лематизован документ у коме само су речи са најчешћим *Universal POS* ознакама (*ADJ*, *NOUN*, *NPROP*, *ADV*, *VERB*, *AUX*, *NUM*, *SYM*, *X*) маскиране тим ознакама;
2. Комбиновањем удаљености докумената добијених у претходном кораку (на нивоу матрица) у нове матрице засноване на композицији, користећи пет различитих метода (Пример 6), и њихово тестирање једне против других као и у претходном кораку;
3. Креирањем перцептрона који користе матрице из првог корака као улазе и њиховим обучавањем на задатку потврђивања ауторства на независном скупу података, како би се обезбедиле одговарајуће и непристрасне тежине за сваку репрезентацију у композитној архитектури;
4. Креирањем и тестирањем композитних матрица без коришћења матрице засноване на *XLM* BERT-u како би се измерио његов утицај на коначне резултате.

6.4.1 Корпус и репрезентације докумената

Корпус над којим је извршен експеримент настао је у оквиру програма *Horizon 2020, COST* акције *Distant Reading for European Literary History*[6] (*Удаљено читање за историју европске књижевности*), у оквиру којег је координисана креација вишејезичног књижевног корпуса *European Literary Text Collection* (Schöch, et al., 2021), *Колекција текстова европске књижевности* (у даљем тексту *ELTeC*), који се састоји од романа писаних на европским језицима иницијално објављених у периоду од 1840. до 1920. године. Прикупљено је по сто романа за једанаест европских језика, а тежило се да романи буду подељени у једнаке групе према полу аутора, броју издања, популарности и временском периоду настанка ([1840–1859], [1860–1879], [1880–1899] и [1900–1920]). Осим тога, десет колекција је анотирано лингвистичком анотацијом

[6] https://www.distant-reading.net, приступљено 31. априла 2022.

укључујући леме, врсте речи, а опционо граматичким категоријама и именованим ентитетима. Овај корпус је у целовитости јавно доступан[7].

У оквиру овог корпуса и српски језик има сопствену колекцију старих романа названу *СрпЕЛТеК* (Krstev, 2021), од којих је укупно сто одабрано овом приликом, за потребе тестирања композитних метода на задатку одређивања ауторства. Свих сто романа (као што је и потребно за генерисање предвиђених репрезентација докумената) анотирано је лемама и врстама речи према *Universal POS* скупу ознака (Stanković, et al., 2021). Колекцију чини укупно 5.861.863 токена, од којих су 4.794.091 речи. Сви романи су подељени у једнаке сегменте од по 10.000 токена, што је резултовало са укупно 544 сегмената које је написало 66 различитих аутора. Од сваког добијеног сегмента креирано је четири различите репрезентације: оригинални документ, лематизован документ, документ триграма ознака врсте речи (*Universal POS* ознаке) и документ биграма креиран од лематизованог текста у коме су најчешће врсте речи замењене ознакама.

6.4.2 Произведене матрице удаљености докумената

Користећи најпре претходно приказану схему (Илустрација 20), припремљено је пет основних матрица удаљености: четири помоћу *Stylo* пакета и листа најфреквентнијих токена према приказаном алгоритму за израчунавање удаљености (Илустрација 18) и једна заснована на векторизацији докумената помоћу вишејезичног *BERT* модела (Табела 9). Број најфреквентнијих токена чије се фреквенције користе као карактеристике за сваку од матрица, као и коришћење н-грама за матрице 3 и 4, предложили су аутори софтвера *Stylo*.

Табела 9: Матрице удаљености докумената добијене коришћењем *Stylo* пакета за *R* и *XLM BERT* модела.

РБ	Назив матрице	Порекло вектора докумената	Метод рачунања удаљености
1	$D_{речи}$	фреквенције 800 најфреквентнијих речи	косинусна *делта* удаљеност
2	$D_{леме}$	фреквенције 800 најфреквентнијих лема	косинусна *делта* удаљеност
3	$D_{ознаке}$	фреквенције 300 најфреквентнијих триграма ознака врста речи	косинусна *делта* удаљеност
4	$D_{маске}$	фреквенције 500 најфреквентнијих биграма речи маскираних врстама речи	косинусна *делта* удаљеност
5	D_{BERT}	*XLM BERT*, према (Iyer & Vosoughi, 2020) и приказаној схеми (Илустрација 19)	косинусна удаљеност

[7] https://zenodo.org/communities/eltec, приступљено 31. априла 2022.

Додатних десет матрица креирано је коришћењем представљених метода комбинације (Пример 6), где је за сваки од пет метода креирано по две матрице, једна која је користила $\boldsymbol{D_{BERT}}$ као део композиције и једна која није (Табела 10).

Табела 10: Матрице удаљености докумената добијене помоћу различитих метода композиције матрица представљених у Табела 9.

РБ	Назив матрице	Начин комбиновања матрица	Комбиноване матрице
6	$D_{просек}$	Аритметичка средина тј. *просечна* вредност свих матрица које се комбинују	*1, 2, 3, 4*
7	$D_{просек\_B}$		*1, 2, 3, 4, 5*
8	$D_{производ}$	*Производ* свих матрица које се комбинују	*1, 2, 3, 4*
9	$D_{производ\_B}$		*1, 2, 3, 4, 5*
10	$D_{минимум}$	Вредност сваког поља је одређена као *минимум* низа вредности свих поља у истом реду и колони у матрицама које се комбинују	*1, 2, 3, 4*
11	$D_{минимум\_B}$		*1, 2, 3, 4, 5*
12	$D_{максимум}$	Вредност сваког поља је одређена као *максимум* низа вредности свих поља у истом реду и колони у матрицама које се комбинују	*1, 2, 3, 4*
13	$D_{максимум\_B}$		*1, 2, 3, 4, 5*
14	$D_{внорма}$	Вредност сваког поља је одређена као *векторска* (l^2) *норма* низа вредности свих поља у истом реду и колони у матрицама које се комбинују	*1, 2, 3, 4*
15	$D_{внорма\_B}$		*1, 2, 3, 4, 5*

Последња група матрица креирана је узимањем аритметичке средине наведених матрица (Табела 9), након што су помножене одговарајућим тежинама према формули:

$$D_{w_B} = \frac{D_{\text{речи}} \cdot w_{\text{речи}} + D_{\text{леме}} \cdot w_{\text{леме}} + D_{\text{ознаке}} \cdot w_{\text{ознаке}} + D_{\text{маске}} \cdot w_{\text{маске}} + D_{BERT} \cdot w_{BERT}}{5}$$

или у случају изузимања D_{BERT} матрице:

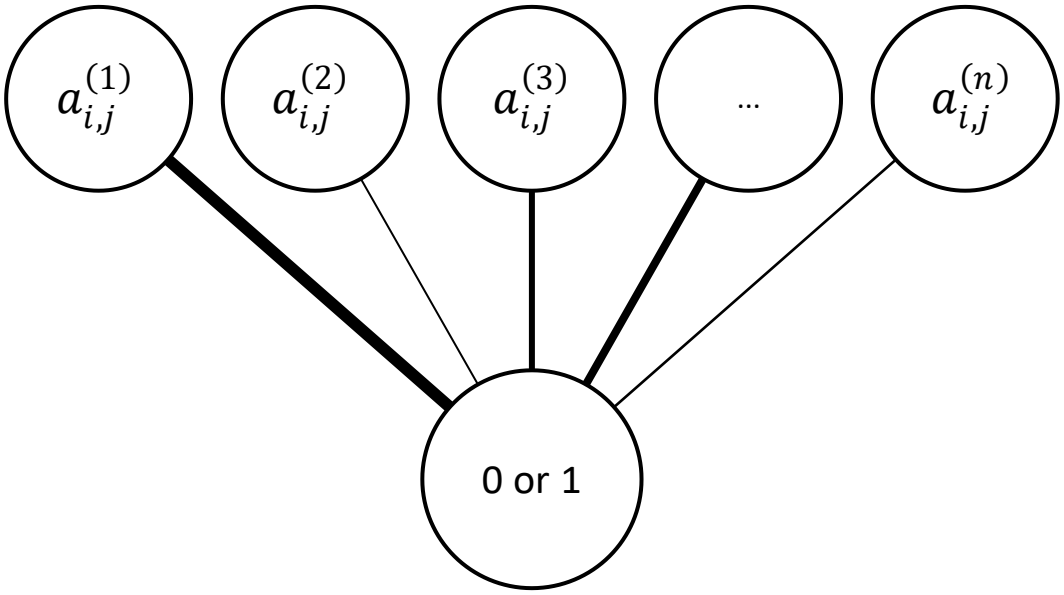
$$D_{w} = \frac{D_{\text{речи}} \cdot w_{\text{речи}} + D_{\text{леме}} \cdot w_{\text{леме}} + D_{\text{ознаке}} \cdot w_{\text{ознаке}} + D_{\text{маске}} \cdot w_{\text{маске}}}{4}$$

при чему је вођено рачуна да се све тежине $w^{(1)}, w^{(2)}, \dots, w^{(n)}$ сумирају тачно до један.

Одговарајуће тежине за сваку од матрица добијене су њиховим обучавањем у оквиру плитке ВНМ, једног перцептрона, где је број улаза једнак броју матрица за које тражимо тежине. Како би направили пресликавање један према један између ивица мреже и тежина које су нам потребне, на излазном слоју налази се само један неурон. Будући да није могуће обучавати перцептрон са једним излазним неуроном за вишекласну класификацију (а имамо више од два могућа аутора) он је уместо тога

обучаван на задатку потврђивања ауторства. Обучавање се обављало на следећи начин (Илустрација 22):

1. Насумично су одабрана два документа (сегмента) који нису из истог романа;
2. За све улазне матрице се добављају вредности удаљености између та два документа;
3. Зарад једноставности, вредности удаљености се инвертују како би се добиле *сличности;*
4. Перцептрон се оптимизује тако да излаз буде 1—уколико два документа потичу од истог аутора или 0—уколико потичу од два различита аутора.



Илустрација 22: Схема за обучавање пожељних тежина за одређене матрице удаљености на задатку потврђивања ауторства.

Ово проузрокује да се тежине које се односе на одређене матрице повећавају уколико су вредности преузете из те матрице добри дискриминатори при утврђивању ауторства. Како би се избегла пристрасност у евалуацији касније, тежине нису обучаване на документима на српском језику, већ на корпусима других шест језика из корпуса *ELTeC* за које је рађен исти експеримент (Škorić, et al., 2022): немачком, енглеском, француском, мађарском, португалском и словеначком.

За сва обучавања коришћен је оптимизатор *ADAM* (Kingma & Ba, 2014) са *почетном стопом учења* (*initial learning rate*) од 0.01, *величином серија (batch size)* фиксираном на 64, а обучавања су трајала 356 епоха. Такође, број парова докумената који су коришћени за обучавање је био балансиран тако да приближно пола парова хетерогеног ауторства, како би се избегла пристрасност ка одређеном избору, а био је балансиран и број докумената који припадају сваком од шест језика.

Обучавањем је добијено два сета универзалних тежина (један укључујући w_{BERT} и један без њега), које се могу применити на српски језик без нарушавања непристрасности (Табела 11).

Табела 11: Тежине које одговарају специфичним приступима у анализи ауторства, добијене обучавањем на задатку потврђивања ауторства за шест европских језика (без српског).

	$w_{\text{речи}}$	$w_{\text{леме}}$	$w_{\text{ознаке}}$	$w_{\text{маске}}$	w_{BERT}
без w_{BERT}	0.507	0.095	0.161	0.236	
са w_{BERT}	0.516	0.052	0.150	0.326	-0.044

Међу њима се одмах може приметити ниска тежина w_{BERT}, што је показатељ да је предвиђени метод слаб дискриминатор стила аутора, бар за српски језик. Такође, може се видети и да највећу тежину носе речи, што је у складу са тренутним резултатима истраживањима у области стилометрије.

Добијене тежине су коришћене за решавање претходно поменутих формула за израчунавање D_w и D_{w_B}, и добијене су две последње матрице (Табела 12).

Табела 12: Матрице удаљености докумената добијене тежински-заснованом комбинацијом матрица представљених у Табела 9.

РБ	**Назив матрице**	**Начин комбиновања матрица**	**Комбиноване матрице**
16	D_w	Пондерисана средина матрица добијена множењем одговарајућим тежинама (добијеним обучавањем перцептрона на задатку потврђивања ауторства)	*1, 2, 3, 4*
17	D_{w_B}		*1, 2, 3, 4, 5*

6.4.3 Евалуација матрица на задатку одређивања ауторства

Над свих седамнаест матрица удаљености сегмената романа, обављена је надгледана евалуација. Нису узете у обзир удаљености између сегмената из истих романа (како би се избегли лаки погоци), а из евалуације су изостављени аутори који нису представљени са бар два романа, како би се постигла атрибуција затвореног типа (јер њихови сегменти нису могли ни са чим бити упарени).

За сваки сегмент који је остао кандидат за евалуацију пронађен је најближи сусед (у виду другог могућег кандидата са најмањом удаљеношћу) како би се креирао скуп парова за евалуацију. Као један тачан одговор се рачунао сваки пар где су оба сегмента уистину креирана од стране истог аутора, док би се као погрешан одговор рачунао сваки пар где сегменти потичу од различитих аутора. На тај начин су за сваку матрицу израчунати тачност, прецизност, одзив и F-мера.

Како бисмо на најбољи начин увидели могуће побољшање перформанси који носи композициони приступ за основицу поређења изабран је резултат најбољег основног метода (што ће се, са српски језик, испоставити да је листа најфреквентнијих речи). Добијени резултати приказани су испод (Табела 13):

Табела 13: Резултати (тачност, прецизност, одзив и F-мера) одређивања ауторства над *СрпЕЛТеК* корпусом, груписана по методу израчунавања удаљености између докумената, где су горњих пет метода самостални, а остали добијени композитним израчунавањем.

РБ	Назив матрице	тачност	прецизност	одзив	F-мера
1	$\boldsymbol{D_{речи}}$	**0.7279**	0.8941	**0.7279**	**0.7518**
2	$D_{леме}$	**0.7279**	0.8312	**0.7279**	0.7364
3	$D_{ознаке}$	0.7082	**0.8973**	0.7082	0.7414
4	$D_{маске}$	0.7016	0.7880	0.7016	0.7140
5	D_{BERT}	0.4918	0.6537	0.4918	0.5226
6	$D_{просек}$	0.7967	**0.9692**	0.7967	0.8120
7	$D_{просек\_B}$	**0.8000**	**0.9692**	**0.8000**	0.8174
8	$D_{производ}$	**0.8000**	0.9668	**0.8000**	0.8160
9	$D_{производ\_B}$	0.7836	0.9406	0.7836	0.8042
10	$D_{минимум}$	0.7344	0.8360	0.7344	0.7388
11	$D_{минимум\_B}$	0.4918	0.6537	0.4918	0.5226
12	$D_{максимум}$	0.7541	0.8896	0.7541	0.7740
13	$D_{максимум\_B}$	0.7541	0.8896	0.7541	0.7740
14	$D_{внорма}$	0.7869	0.9622	0.7869	0.8067
15	$D_{внорма\_B}$	0.7869	0.9622	0.7869	0.8067
16	D_{w}	0.7934	0.9655	0.7934	**0.8169**
17	D_{w_B}	0.7934	0.9673	0.7934	**0.8181**

Резултати тачности, прецизности, одзива и Ф-мере за свих седамнаест матрица удаљености су потом додатно рекомпиловани тако да приказују проценат побољшања композитних метода у односу на основицу (*baseline*), израчунавање удаљености засновано на листи најфреквентнијих речи. Ови резултати приказани су у наставку текста (Табела 14).

Табела 14: Проценат унапређења према свакој мери који су композитни методи остварили у односу на основицу (засновану на листи најфреквентнијих речи).

РБ	Назив матрице	тачност	прецизност	одзив	F-мера
6	$D_{просек}$	9.45%	8.40%	9.45%	8.01%
7	$D_{просек\_B}$	**9.91%**	8.40%	**9.91%**	8.73%
8	$D_{производ}$	**9.91%**	8.13%	**9.91%**	2.95%
9	$D_{производ\_B}$	7.65%	5.20%	7.65%	2.95%
10	$D_{минимум}$	0.89%	-6.50%	0.89%	-1.73%
11	$D_{минимум\_B}$	-32.44%	-26.89%	-32.44%	-30.49%
12	$D_{максимум}$	3.60%	-0.50%	3.60%	8.54%
13	$D_{максимум\_B}$	3.60%	-0.50%	3.60%	6.97%
14	$D_{внорма}$	8.11%	7.62%	8.11%	7.30%
15	$D_{внорма\_B}$	8.11%	7.62%	8.11%	7.30%
16	D_{w}	9.45%	8.59%	9.45%	**9.34%**
17	D_{w_B}	8.11%	5.45%	8.11%	6.89%

Из приложених резултата закључујемо да је оптималан начин креирања матрице удаљености за потребе утврђивања ауторства у овом случају коришћењем листе најфреквентнијих речи што се тиче самосталних метода, док је свеукупно најбољи метод множења матрица (што се тиче тачности) и тежински-подржано усредњавање (комбиновање засновано на обучавању перцептрона), што се тиче Ф-мере. За ово унапређење, коришћењем *Њукомб-Вилсоновог* теста, израчуната је и статистичка значајност која износи p = 0.04526, што је значајно за стандардни ниво од α = 0.05, па тако несумњиво добијамо потврдан одговор и на ИП3:

Да ли паралелни језички модели боље моделирају језик од појединачних?

Најиздвојеније резултате (у негативном смислу) постиже употреба минимизирања вредности матрица, поготово са коришћењем трансформера *BERT*, где се резултат смањио за преко 30% у свим категоријама. Осим тога, композитни методи су умногоме надмашили стандард у пољу (унапређење тачности и Ф-мере од скоро 9%), што додатно потврђује претходни одговор на ИП2:

Да ли се композитни интелигентни системи засновани на паралелним моделима могу надограђивати – и да ли већи број активних паралелних модела побољшава квалитет целокупног система?

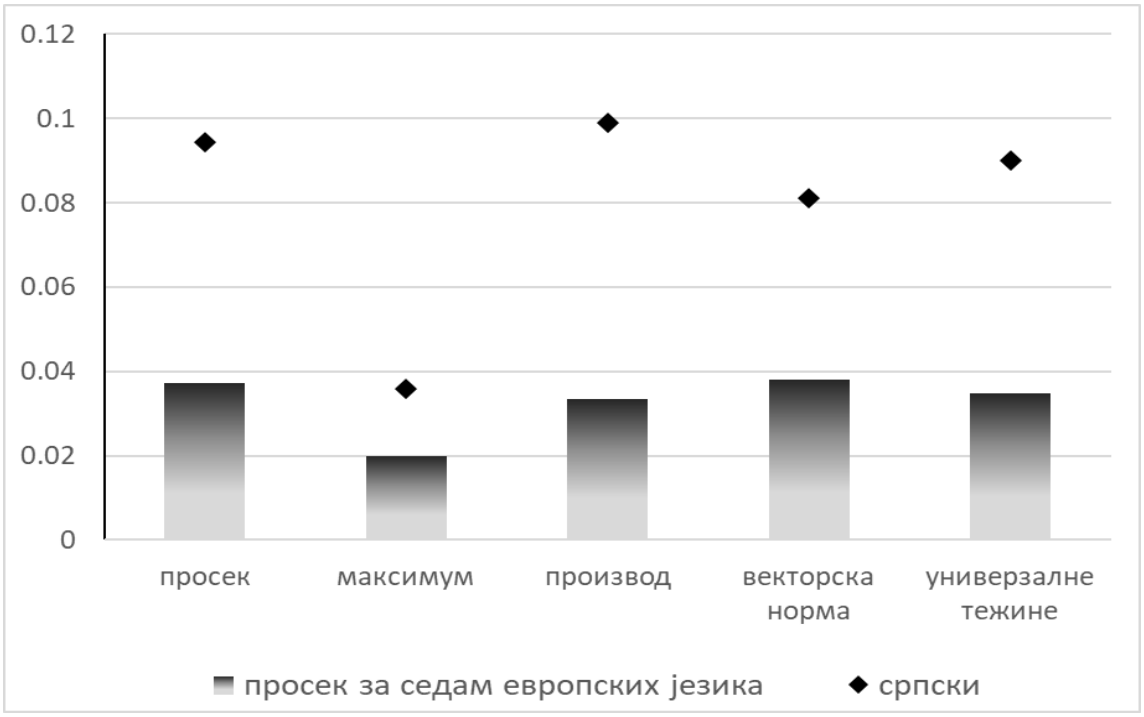
Ипак, остаје неразјашњено да ли је наслагани класификатор најбољи метод комбиновања (упркос највишој оствареној Ф-мери), јер су разлике између најбољих

комбинованих приступа у овом случају статистички незначајне, па тако, тренутно, остаје непознат одговор на ИП5:

Који су оптимални методи комбинације излаза појединачних језичких модела?

Што се тиче употреба трансформера *BERT* за векторизацију докумената у циљу одређивања ауторства у овом случају нису пронађени докази да је то пожељно. Уколико изузмемо резултате остварене од стране $D_{минимум\_B}$, у просеку, добија се унапређење од 0.07% коришћењем композиција које садрже D_{BERT} матрицу, што је ипак статистички безначајно.

Дакле, из резултата овог експеримента закључујемо да је употреба паралелних матрица удаљености у моделирању стила аутора адекватан метод који доноси велика унапређења. Даље, установљено је неколико различитих метода комбинација који производе добро резултате и који се даље могу истражити на пољу општег моделирања природног језика, што се намеће као следећи корак у истраживању ове методе. Оно што је посебно занимљиво из перспективе овог рада је то да су добијена унапређења од додатног значаја за моделирање српског језика, где су унапређења (добијена од стране свих композитних метода) у просеку знатно већа у односу на она добијена за друге европске језике (Илустрација 23).



Илустрација 23: Унапређења резултата за задатак одређивања ауторства која се добијају коришћењем паралелних архитектура за српски језик (квадратић), упоређена са просечним унапређењима која се добијају за седам европских језика (стубови) на истом експерименту.

III Композитне псеудограматике српског језика

Поставка

У претходним одељцима успешно је (потврдно) одговорено на три истраживачка питања:

- **ИП1** Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и њиховим излазним вероватноћама адекватна метода у области обраде природних језика? (одељак 5)
- **ИП2** Да ли се композитни интелигентни системи засновани на паралелним моделима могу надограђивати – и да ли већи број активних паралелних модела побољшава квалитет целокупног система? (одељци 5 и 6)
- **ИП3** Да ли паралелни језички модели боље моделирају језик од појединачних? (одељак 6)

и још увек је потребно одговорити на преостала три:

- **ИП4** Да ли композитне псеудограматике боље моделирају језик од појединачних?
- **ИП5** Који су оптимални методи комбинације излаза појединачних језичких модела?
- **ИП6** Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и излазним вероватноћама подобна метода у области интелигентних система уопште?

Са тиме на уму извршен је додатни експеримент приликом којег је развијено неколико јединствених језичких модела (више о томе у одељку 7), који су потом на различите начине комбиновани како би се реализовале различите псеудограматике српског језика (одељак 8). Уз то, реализована је и њихова евалуација на седам различитих задатака (одељак 9), са циљем добијања дефинитивних одговора на ИП4 и ИП5, па самим тим и на ИП6.

7
Припрема језичких модела и других ресурса

Обучавање језичких модела који би послужили као основа за одговор на преостала истраживачка питања, састојало се из три корака:

1. Прикупљање корпуса текстова. Како би се истраживање односило првенствено на српски језик, било је неопходно припремити адекватан корпус за обучавање и тестирање модела;
2. Трансформација корпуса у различите репрезентације. Прикупљени корпус било би пожељно додатно трансформисати. Ранија истраживања су показала да обрађивање линеарног текста даје знатно слабије резултате у односу на текст који је додатно трансформисан тј. проширен (анотиран) класама атрибута: врсте речи, функције, припадности различитим класама и/или доменима. На основу различитих атрибута (било постојећих или накнадно изведених) прикупљени корпус је трансформисан на различите начине, како би се на њему обучили различити језички модели;
3. Обучавање језичких модела. На свакој од припремљених репрезентација корпуса обучен је јединствени језички модел заснован на архитектури генеративног предобученог трансформера друге генерације (ГПТ-2), а за потребе различитих репрезентација припремљени су и специфични алати за адекватну предобраду и токенизацију;

7.1 Припрема корпуса

Као корпусна основа за ово истраживање понајвише су послужиле верзије Корпуса савременог српског језика, *СрпКор2013*[8], и његова најновија надоградња, *СрпКор2021*[9], које су развили истраживачи Универзитета у Београду и Друштва за језичке ресурсе и технологије (JePTex)[10]. Друштво за језичке ресурсе и технологије, које окупља истраживаче са неколико факултета Универзитета у Београду, Института за српски језик САНУ и других институција, највећи је произвођач (генератор) језичких ресурса и алата за обраду српског језика (Krstev & Stanković, 2022).

7.1.1 Корпус савременог српског језика – СрпКор2013

Група за језичке технологије Универзитета у Београду (*ОПЈ група*) Математичког факултета Универзитета у Београду под руководством проф. др Душка Витаса припремила је прву, неанотирану верзију Корпуса савременог српског језика, *СрпКор2003*, доступну на веб презентацији Математичког факултета. Корпус *СрпКор2003*, величине 22.2 милиона речи, чине новински чланци објављени после 1994. године, одабрани уџбеници и монографије објављени после 1981. године и литература објављена после 1920. године (Krstev & Vitas, 2005).

У наредних десет година (а поготову у периоду од 2009. до 2013. године) на проширењу корпуса радили су чланови и сарадници ОПЈ групе, заједно са студентима Филолошког, Математичког и Рударско-геолошког факултета. Иако је званично објављен 2013. године (Utvić, 2014), ресурс се води као динамички јер се и даље допуњује. У њему преовлађују текстови оригинално писани на српском језику (92.74% текстова и 90.47% токена), објављени после 2000. године (87% текстова и 89% токена) и то претежно новински (66.4% текстова и 73.7% корпусних речи), док 14.8% речи потиче из књижевноуметничких, научних и научно-популарних текстова, а 5.6% речи потиче из текстова административног стила (Утвић, 2013). Анотиран је ознакама врсте речи и лемама помоћу тагера *TreeTagger*, и доступан је у оквиру веб презентација математичког факултета[8] и *NoSketch* апликације (Kilgarriff, et al., 2014) друштва JePTex[9] уз посебну регистрацију. СрпКор2013 је у својој целости (145.3 милиона токена и 121.2 милиона речи) коришћен за обучавање ГПТ-2 модела креираних за потребе овог истраживања.

7.1.2 Проширење Корпуса савременог српског језика - СрпКор2021

Крајем 2021. године објављена је велика допуна Корпуса савременог српског језика под називом СрпКор2021. Верзија СрпКор2021 садржи више од 700 милиона токена и више од 600 милиона речи, и аутоматски је обележена ознакама врсте речи и лемама применом композитног тагера претходно описаног у одељку 5. Нови текстови

[8] http://www.korpus.matf.bg.ac.rs, приступљено 1. јула 2022.

[9] http://noske.jerteh.rs, приступљено 1. јула 2022.

[10] http://jerteh.rs, приступљено 1. јула 2022.

већином воде порекло од новинских чланака са веб портала, међутим, укључене су и докторске дисертације, романи, законска регулатива (и други текстови из домена права), стручна литература (и уџбеници), као и текстови настали транскрипцијом говорног језика, претежно транскрипцијом седница Народне скупштине Србије и радио емисија.

Сви текстови су припремљени на исти начин као и они за СрпКор2013: новински чланци су прикупљени током дужег периода повременим *гребањем веба* уз процедуре прилагођене структурама појединих информативних портала, а текстови су додатно очишћени од сувишних етикета, линкова, рекламних делова и коментара. Део чланака је додатно ручно коригован, при чему су исправљане и штампарске грешке. Литерарни део корпуса је мањим делом преузет из постојећих електронских извора, као што су *Амстердамски словенски паралелни поравнати корпус* (*Amsterdam Slavic Parallel Aligned Corpus*, *ASPAC)*[11] и *Антологија српске књижевности* (АСК)[12], а већим делом је настао дигитализацијом која је почела скенирањем, након чега је рађено оптичко препознавање карактера, аутоматска провера и обележавање текста (Krstev & Stanković, 2020) и коначно, ручно кориговање. Докторске дисертације Универзитета у Србији преузете су из *Националног репозиторијума дисертација Србије* (НаРДуС)[13], при чему је селекција урађена ручно, а ручно су кориговани и одређени делови дисертација. За монографије, дисертације и друге садржаје који су били доступни у ПДФ формату, поступак је био сличан, при чему су први кораци (сканирање и оптичко препознавање) изостављени као непотребни у случају неких дигитално-рођених докумената са електронским текстом.

Ослањајући се на морфолошке речнике (Krstev, 2008; Vitas & Krstev, 2012), урађена је и глобална провера за све текстове који се припремају за корпус, укључујући пребројавање препознатих и непрепознатих речи, интерпункцијских знакова и слично, на основу чега је грубо оцењена процентуална *коректност* текста. На основу израчунатих коректности свих текстова, као и њиховог типа, одабрани су и они који ће бити коришћени за потребе овог рада. Пробрани су, дакле, само литерарни, научни и новински текстови коректности преко 98%, што је резултовало поткорпусом од приближно 58.5 милиона токена (око 7% укупне величине), док су остали текстови враћени на још једну итерацију полуаутоматске провере.

7.1.3 Остали корпуси

Осим поменуте две верзије СрпКор-а, при обучавању су коришћена још два корпуса, *ВикиКорпус* (81.3 милиона токена) и *СрпЕЛТеК* (5.9 милиона токена) (Krstev & Stanković, 2022), доступна у отвореном приступу путем *NoSketch* апликације друштва JeРТех[9]. ВикиКорпус садржи текстове са Википедије на српском језику, прикупљене аутоматским гребањем, који су потом полуаутоматски прочишћени, док се СрпЕЛТек

[11] https://spraakbanken.gu.se/en/resources/aspacsvsbc, приступљено 1. јула 2022.

[12] http://www.antologijasrpskeknjizevnosti.rs, приступљено 1. јула 2022.

[13] https://nardus.mpn.gov.rs/, приступљено 1. јула 2022.

састоји од 100 романа који чине српску потколекцију корпуса *ELTeC* (колекције текстова европске књижевности), припремљену у оквиру *COST* акције *CA16204 Distant Reading for European Literary History* (Удаљено читање за историју европске књижевности), описаног у одељку 6.4.1.

Коришћењем сва четири поменута извора (СрпКор2013, део СрпКор2021, ВикиКорпус и СрпЕЛТеК), створен је корпус величине приближно 291 милиона токена, који ће у даљем тексту бити називан *основни корпус за обучавање* (Табела 15).

Табела 15: Порекло и структура основног корпуса за обучавање језичких модела коришћених за ово истраживање.

Назив корпуса	тип	величина (у милионима токена)	величина исечка за нови корпус	постотак удела у новом корпусу
СрпКор2013	мешовити	145.2	145.2	50%
СрпКор2021	мешовити	716.8	58.5	20%
ВикиКорпус	општи	81.3	81.3	28%
СрпЕЛТеК	литерарни	5.9	5.9	2%

7.2 Проширење корпуса кроз различитим репрезентацијама

Као што је већ поменуто, креирање самосталних језичких модела за потребе овог истраживања ослањаће се на различите, специјално припремљене репрезентације прикупљеног корпуса. У циљу стварања самосталних модела који представљају различите аспекте језика (и самим тим производе потенцијално различите перплексности када су суочени са истим исечком текста) осмишљене су две репрезентације, које представљају два комплементарна аспекта текста на природном језику: семантику и синтаксу текста.

Треба нагласити да се у овом раду под семантиком текста подразумевају значења одељака тог текста, екстрахована методама *латентне семантичке анализе* (Evangelopoulos, 2013), док се синтакса односи на граматичка правила екстрахована методама корпусне лингвистике. Са тим на уму, креирана су два алгоритма за екстракцију информација из текста (један за семантичку и један за синтаксичку екстракцију) којима ће се креирати различите репрезентације текста (и самим тим проширити корпус за обучавање), и који ће се уједно користити за предобраду улаза одговарајућих језичких модела. Такође, како би се извршиле поменуте трансформације текста у различите репрезентације, текст мора бити анотиран одговарајућим додатним информацијама, па је било неопходно припремити ресурсе и функције и за тај аспект обраде.

7.2.1 Семантичка репрезентација текста

Креирање семантичких репрезентација текста за потребе овог истраживања ослања се, као што је већ поменуто, на методе латентне семантичке анализе, прецизније, на смањење димензионалности текста кроз уклањање токена који не преносе значење (стоп-речи) и пресликавања преосталих токена у класе. Овај процес се извршава у три корака (Илустрација 24, Пример 7):

Илустрација 24: Алгоритам трансформације текста у његову семантичку репрезентацију.

1. Анотација текста. У овом случају, анотација се користи за допуњавање текста неопходним информацијама које би омогућиле потребну обраду, а те информације су ознаке врсте речи (укључујући интерпункцију као засебну ознаку) адекватних токена и њихове леме. Анотација се обавља коришћењем софтвера описаног у одељку 5 (Stanković, et al., 2022), а скуп ознака који се користи је *Universal POS*, док су за лематизацију коришћени морфолошки речници српског језика (Krstev, 2008);
2. Елиминација интерпункције и стоп-речи. Елиминација се врши на основу ознака добијених у претходном кораку, при чему се из текста елиминишу сви токени осим оних који су означени као именице (*NOUN* или *PROPN*), придеви (*ADJ*), бројеви (*NUM*), глаголи изузев помоћних (*VERB*) и прилози (*ADV*), у циљу очувања само токена који су превасходни носиоци значења;

3. Пресликавање преосталих токена у класе. У овом кораку се преостали текст лематизује, тј. преостали токени се замењују њиховим лемама, добијеним анотацијом у првом кораку.

Изворни текст	*Ако се задесиш у Риму, понашај се као Римљанин. Ако се задесиш другде, понашај се према тамошњим обичајима.*
Текст обележен ознакама врсте речи и лемама	*Ако[SCONJ, ако]* *се[AUX, се]* *задесиш[**VERB**, задесити]* *у[ADP, у]* *Риму[**PROPN**, Рим]* *,[PUNCT, ,]* *понашај[**VERB**, понашати]* *се[AUX, се]* *као[CCONJ, као]* *Римљанин[**PROPN**, Римљанин]* *.[PUNCT, .]* *Ако[SCONJ, ако]* *се[AUX, се]* *задесиш[**VERB**, задесити]* *другде[**ADV**, другде]* *,[PUNCT, ,]* *понашај[**VERB**, понашати]* *се[AUX, се]* *према[ADP, према]* *тамошњим[**ADV**, тамошњи]* *обичајима[**NOUN**, обичај]* *.[PUNCT, .]*
Текст у коме су сачуване само именице, придеви, бројеви, прилози и глаголи (изузев помоћних)	*задесиш[VERB, **задесити**]* *Риму[PROPN, **Рим**]* *понашај[VERB, **понашати**]* *Римљанин[PROPN, **Римљанин**]* *задесиш[VERB, **задесити**]* *другде[ADV, **другде**]* *понашај[VERB, **понашати**]* *тамошњим[ADV, **тамошњи**]* *обичајима[NOUN, **обичај**]*
Семантичка репрезентација текста	*задесити Рим понашати Римљанин задесити другде понашати тамошњи обичај*

Пример 7: Трансформација текста у семантичку репрезентацију по корацима на примеру познате изреке.

7.2.2 Синтаксичка репрезентација текста

Синтаксичка репрезентација текста је за потребе овог рада осмишљена као комплементарна претходно утврђеној семантичкој репрезентацији, те је циљ при креирању таквих репрезентација био да се оне максимално лише значења, али да остану граматички исправне. То је постигнуто, укратко, заменом токена који преносе значење (оних који се чувају за семантичку репрезентацију) њиховим надређеним граматичким категоријама, што би, у смислу теорије формалних граматика, имало исти ефекат као кад би се одређене граматичке категорије (иначе нетерминали) уврстили у алфабет граматике (скуп терминала).

Како саме ознаке врста речи (а поготово оних променљивих) у појединим случајевима нису довољне за опис граматичких правила на српском језику (на пример, именичке синтагме типа *придев* + *именица* у којима се захтева слагање именице и придева у роду, броју и падежу, или пак предлошко-падежне конструкције типа *предлог* + *именица* у којима одређени предлог захтева именицу у одређеном падежу), било је неопходно да се поједини делови текста, односно специфични токени, обележе додатним граматичким категоријама.

Коначно решење за креирање синтаксичке репрезентације текста огледа се у чувању свих непроменљивих речи (сем прилога, који су замењени ознаком за прилог *ADV*), заменица и интерпункције у изворном облику (који је једнак леми), док се променљиве речи осим заменица (именице, придеви, бројеви, глаголи) замењују адекватним граматичким описом који, поред врсте речи, садржи и граматички род, број, падеж и аниматност за именице, степен компарације и вид за придеве, те лице и време (облик) за глаголе. Граматичке категорије су реализоване према ознакама за морфосинтаксичко обележавање српског језика (Krstev, et al., 2004; Vitas & Krstev, 2012; Stanković, et al., 2018)[14]. Коришћењем описане обраде се за изворни текст (Пример 7) добија следећа синтаксичка репрезентација (Пример 8):

Ако се задесиш у Риму, понашај се као Римљанин. Ако се задесиш другде, понашај се према тамошњим обичајима.

Ако се ***Pys*** *у* ***fs4q***, ***Yys*** *се као* ***ms1v***. *Ако се* ***Pys ADV***, ***Yys*** *се према* ***aems6g mp6q***.

Пример 8: Текста трансформисан у синтаксичку репрезентацију (доле) према граматичким категоријама (подебљано) на примеру познате изреке (горе).

За потребе овог истраживања, аутоматска трансформација текста у синтаксичку репрезентацију (Пример 8), остварена је помоћу специјално припремљеног модела за *TreeTagger*, обученог на анотираном корпусу српског језика који је настао у оквиру пројекта *MULTEXT-EAST* (Krstev, et al., 2011), са тим да су у скупу за обучавање ознаке непроменљивих речи биле замењене лемама тих речи (сем у случају прилога, где су

[14] http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html, приступљено 1. јула 2022.

све ознаке замењене са *ADV*). Модел тагера за аутоматску трансформацију текста је обучен са стандардним параметрима, а за лексикон је припремљен дериват система морфолошких речника српског (Krstev, 2008). Једном направљен модел је спреман да анотира било који нови текст писан на српском језику у циљу његове трансформације у предвиђену синтаксичку репрезентацију.

7.3 Обучавање језичких модела

Обучавање самосталних језичких модела из класе генеративних предобучених трансформера друге генерације (ГПТ-2) за потребе овог истраживања одвијало се уз коришћење библиотеке *transformers* (Wolf, et al., 2019) за програмски језик Пајтон (*Python*), коју је развила заједница *Hugging Face*[15]. С обзиром на то да обучавање модела коришћењем ове библиотеке захтева припремљен речник токена (Пример 9) и одговарајући предобучени токенизатор, ти ресурси (за српски језик) су преузети са платформе *Hugging Face*[15], при чему је речник (иницијално припремљен за токенизацију ћириличног текста) пажљивом ручном транслитерацијом прилагођен латиничним корпусима припремљеним за обучавање модела и описаним у одељку 7.1.

5	„!“	1691	„jet“
6	„\““	1692	„Ġgener“
7	„#“	1693	„Ġmo“
8	„$“	1694	„oka“
9	„%“	1695	„abr“
…		…	
37	„A“	17955	„inacije“
38	„B“	17956	„mera“
39	„C“	17957	„Ġpaket“
40	„D“	17958	„Ġupoznavanje“
41	„E“	17959	„Ġmajkl“
…		…	
193	„Ā“	51954	„Ġnerad“
194	„ā“	51955	„Ġgornju“
195	„Ă“	51956	„Ġlegura“
196	„ă“	51957	„Ġbratski“
197	„Ą“	51958	„Ġprivatizacija“

Пример 9: Исечци речника токена груписани према редном броју. Карактер Ġ означава белину, или у овом случају, то да одређени токен није афикс већ реч (или започиње реч када је реч о токенима који представљају подречи).

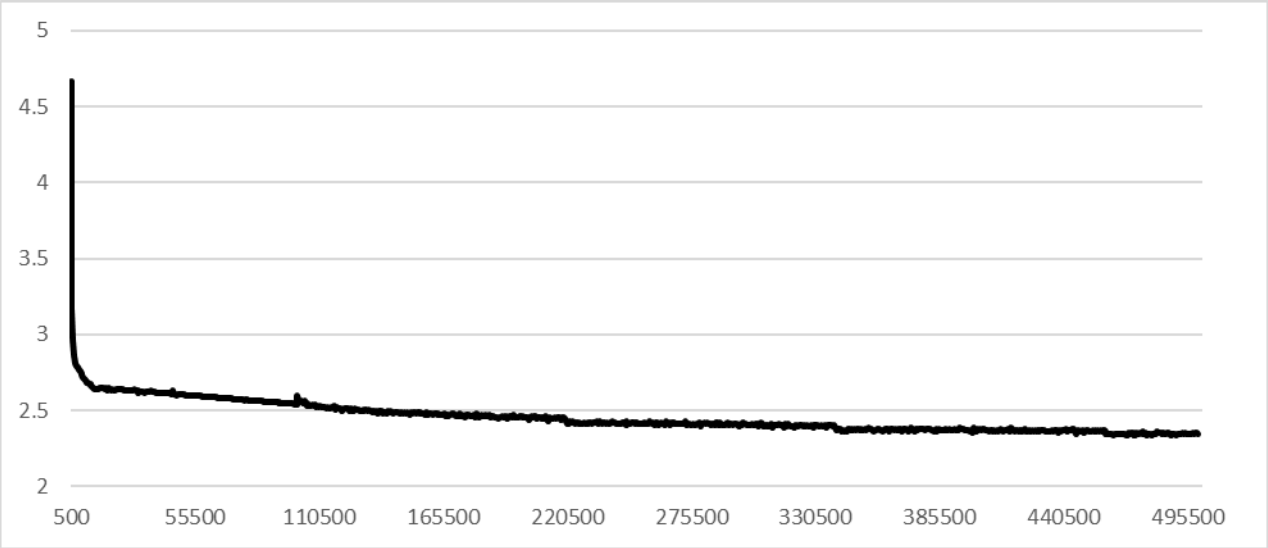
[15] https://huggingface.co, приступљено 1. јула 2022.

Обучена су, дакле, три модела: такозвани *основни*, *семантички* и *синтаксички*, сваки на одговарајућем, припремљеном корпусу, а уз предефинисана подешавања: ГПТ-2 модел од 117 милиона параметара са 12 слојева, 12 глава пажње и векторима величине 768, и уз припремљени поменути латинични речник од 52,000 токена. Табела 16 приказује детаљнија подешавања коришћена при обучавању модела.

Табела 16: Подешавања за обучавање трансформера библиотеке *transformers,* коришћена за потребе овог истраживања.

параметар	**вредност**
activation_function	gelu_new
architectures	[GPT2LMHeadModel]
attn_pdrop	0.1
bos_token_id	0
embd_pdrop	0.1
eos_token_id	2
gradient_checkpointing	FALSE
initializer_range	0.02
layer_norm_epsilon	1.00E-05
model_type	gpt2
n_ctx	1024
n_embd	768
n_head	12
n_inner	null
n_layer	12
n_positions	1024
resid_pdrop	0.1
scale_attn_weights	TRUE
summary_activation	null
summary_first_dropout	0.1
summary_proj_to_labels	TRUE
summary_type	cls_index
summary_use_proj	TRUE
use_cache	TRUE
vocab_size	52000

За обучавање првог од три модела коришћен је основни, неизмењени корпус за обучавање, при чему је корпус најпре подељен на текстове[16], који су насумично промешани и подељени у две неопходне групе: текстови за обучавање (отприлике 90% од укупног броја припремљених реченица) и текстови за валидацију (отприлике 10% реченица). Параметри BHM су насумично подешени. Модел је обучаван у трајању од 500 хиљада корака или 4,17 епоха, при чему је на скупу за валидацију постигнута минимална ентропија (*loss*) у вредности од 2,3328 (Илустрација 25).



Илустрација 25: Крива промене вредности ентропије (*loss*) у односу на скуп за валидацију при обучавању основног модела за потребе овог истраживања.

Обучени основни модел, дакле, има функцију генерисања уобичајеног текста на српском језику (Пример 10), а када је у питању евалуација, модел би оцењивао колика је, уопштено, вероватноћа да задата ниска припада српском језику.

и преко њега прелази у грчки град Крф. Ту је остао до своје смрти, 10. септембра 497. године. Срби славе Светог Ђорђа 27. док Српска

Након завршетка Првог светског рата, вратио се у Југославију где је провео остатак живота као професор Филозофског факултета у Београду, где је завршио и

, јер је пре тога био председник владе у егзилу. Још увек је био истакнути члан КПЈ, али је остао упамћен као један од најистакнутијих чланова КПЈ.

Пример 10: Пример три исечка текста које генерише основни обучени језички модел[17], при чему су за сваки пример коришћени различити замеци (*seed*) и насумично генерисање без улазног текста.

[16] Корпус је подељен на текстове, а не на реченице из разлога што је препоручено да се модели обучавају на што дужим, повезаним, комадима текста.

[17] Приказани текст је пресловљен у ћирилицу.

Оно што је битно напоменути је то да се ниједна приказана ниска (Пример 10) не налази у том облику у корпусу који је коришћен за обучавање модела. Највеће подударање у смислу дужине има ниска

Након завршетка Првог светског рата, вратио се у

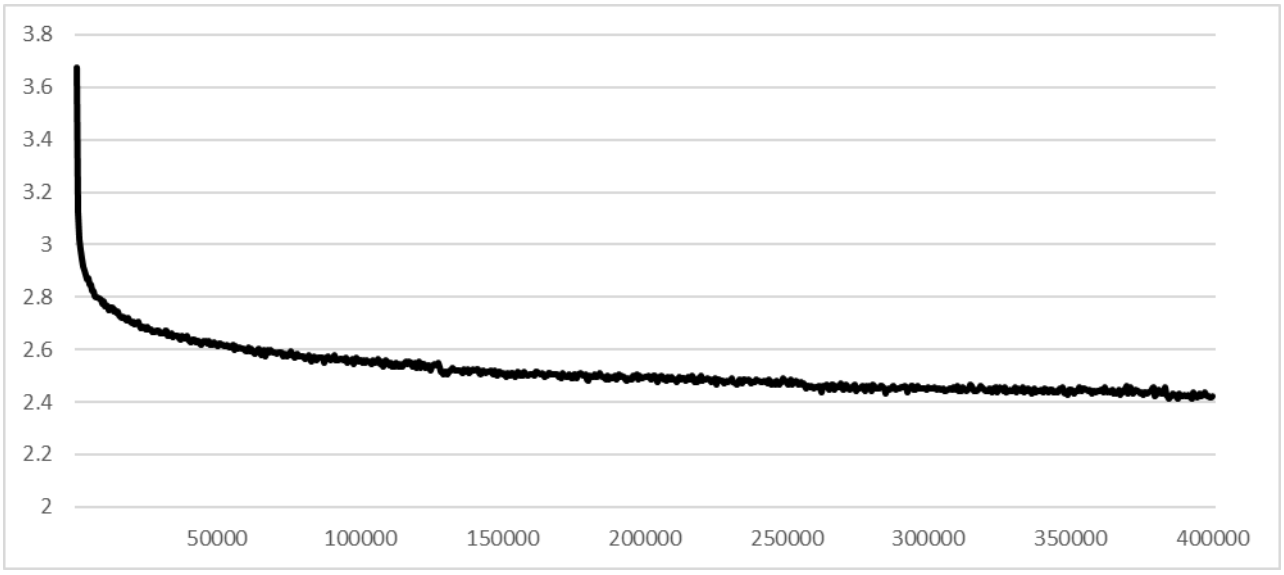
која је део реченице

Након завршетка Првог светског рата, вратио се у Волоско али пошто није било посла, с братом Иваном је отишао у Филах, Аустрија.

Осим поменутог примера у корпусу за обучавање налазе се и подниске:

- *професор Филозофског факултета у Београду* (128 понављања)
- *владе у егзилу* (65 понављања)
- *Још увек је био* (39 понављања)
- *вратио се у Југославију* (35 понављања)
- *један од најистакнутијих чланова* (19 понављања)
- *је остао до своје смрти* (17 понављања)
- *али је остао упамћен као* (2 понављања)

Обучавање другог, *семантичког* модела, обављено је као дообучавање основног модела, при чему није промењен ниједан параметар, већ се променио само скуп текстова (овога пута коришћен је семантички корпус), а скупови за обучавање и валидацију су креирани на исти начин као и за први модел. С обзиром на то да је овога пута задатак обучавања био лакши (услед тога што се обучавање није вршило *од нуле*), оно је трајало 400 хиљада корака или 3,34 епохе, при чему је у односу на скуп за валидацију постигнута минимална ентропија у вредности од 2,4101 (Илустрација 26).



Илустрација 26: Крива промене вредности ентропије (*loss*) у односу на скуп за валидацију при обучавању семантичког модела за потребе овог истраживања.

Овако обучен модел има функцију генерисања ниске текста која није граматички исправна, али ипак *прича неку причу* (Пример 11), те би се од ње уз додатну обраду могао направити смислен текст. Са друге стране, у случају евалуације, овај модел би (у идеалном случају) оцењивао степен смислености задатог текста (трансформисаног у одговарајући облик коришћењем процедуре из одељка 7.2.1).

предвидети дати закон примењивати исти година предвиђати дати предложити закон представљати препрека приближавање

кренути рат остати запамтити ратнички вештина ратовање непријатељски војска приписивати прек војник ратник приказати рат имати ћерка

потписати уговор сарадња привредни комора представник приватан компанија истакати дати представљати први корак приближавање земља унија

Пример 11: Пример три исечка текста које генерише обучени *семантички* језички модел[17].

Обучавање, последњег, *синтаксичког* модела обављено је, поново, као дообучавање првог, основног модела, само уз промену скупова за обучавање и валидацију. Овом приликом је битно напоменути да речник који се користи за токенизацију (Пример 9) улазног (и излазног) текста (при обради од стране језичког модела) садржи и токене дужине један (карактере), у служби склапања нових токена који нису појединачно дефинисани у речнику. Из тог разлога, била је неопходна додатна трансформација, како не би дошло до двосмислености. На пример, ниска *np* означава именицу средњег рода у множини (према коришћеном скупу ознака граматичких категорија[14]), али се не појављује у том смислу у ниски карактера *jedanput*. Исто тако, карактер *S* не означава нужно глаголски прилог за садашње време. Због тога, као и како би се елиминисале друге непредвиђене околности, направљено је пресликавање свих ниски које означавају граматичке категорије у одговарајуће ниске карактера који нису уобичајени за српски језик, а налазе се у поменутом речнику токена (Табела 17).

Табела 17: Пример пресликавања ознака граматичких категорија у недвосмислене ниске карактера.

Ознака категорије	Пресликана ниска	Опис граматичке категорије
Yys	Áôë	Императив другог лица једнине
ms7g	Þëćû	Именица мушког рода, једнина, локатив, анимантност није од значаја
fs7v	ßëćù	Именица женског рода, локатив, аниматна
cemp7g	øòÞìćû	Суперлатив мушког рода, множина, локатив, анимантност није од значаја
fs1g	ßëüû	Именица женског рода, номинатив, анимантност није од значаја
nw4q	êíÿú	Именица средњег рода, паукал, акузатив, аниматна

akms2g	öîÞëýû	Позитив мушког рода, неодређени облик, једнина, генитив, аниматност није од значаја
fw2v	ßíýù	Именица женског рода, паукал, генитив, аниматна
fp5q	ßìăú	Именица женског рода, множина, вокатив, аниматна
cemp3g	øòÞìþû	Суперлатив мушког рода, множина, датив, аниматност није од значаја
ADV	±	Прилог

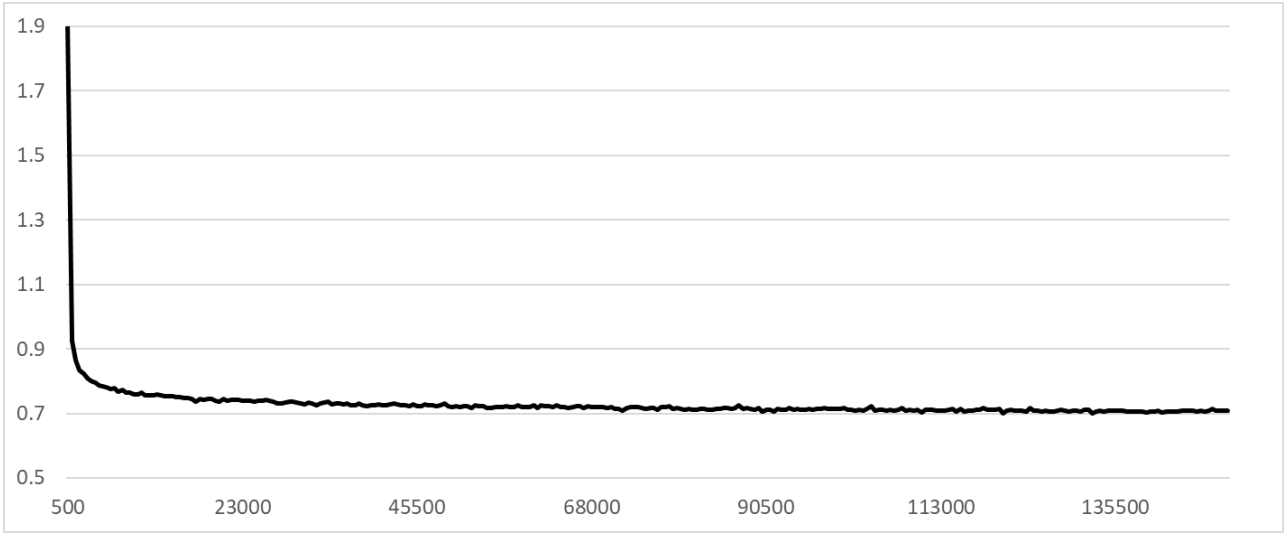
Применом пресликавања би се за семантичку репрезентацију (Пример 8) добио следећи текст (Пример 12):

*ако се **Pys** у **fs4q**, **Yys** се као **ms1v**. ако се **Pys ADV**, **Yys** се према **aems6g mp6q**.*

*ako se **Æôë** u **ßëÿú**, **Áôë** se kao **Þëüù**. ako se **Æôë ±**, **Áôë** se prema **aeÞëąû Þìąú**.*

Пример 12: Трансформација ознака граматичких категорија из синтаксичке репрезентације текста (горе) у ниске карактера који нису део српског језика (доле)[17].

Како је обучавање оваквог модела (у великој мери заснованог на карактерима, јер се граматичке категорије не налазе у речнику, већ се креирају комбинацијама карактера) била непознаница, дужина обучавања је подешена на милион корака. Обучавање се доста разликовало од претходна два случаја, где се вредност ентропије кретала од 1,9 све до минимума од 0,6654 (што је вероватно последица погађања узастопних карактера у приказима граматичких категорија), и при чему је након првих 150 хиљада корака смањење ентропије у односу на скуп за валидацију било минимално (Илустрација 27).



Илустрација 27: Крива промене ентропије у односу на скуп за валидацију при обучавању синтаксичког модела за потребе овог истраживања.

Овако обучен (синтаксички) модел има функцију генерисања ниске текста у којој су одређени токени генерализовани и замењени граматичком категоријом (Пример 13), те када би се заменили правим облицима речи (који припадају тим категоријама) могао би се направити смислен текст. Са друге стране, у случају евалуације, овај модел би (у идеалном случају) оцењивао синтаксичку исправност задатог текста, под условом да је трансформисан у одговарајући облик, коришћењем функције за трансформацију описане у одељку 7.2.2.

Þëýú Þìýù Òëß и ßëüú êë2q aeÞÿû ßìÿú ßíÿëù, односно êì5q. Þë

ßëýú, Æõë Þëÿú Þìýù, Þíÿù Þú, öðßëüû ÞñÞëħù и Òíê ßìăú.

öðßëýû ßëþú, ÒìÞ су да Æõì Þëÿú. само је Þìăú Þíÿù ßìýú Òë

Пример 13: Пример три исечка текста које генерише обучени *синтаксички* језички модел[17].

Обучавање свих језичких модела поменутих у овом поглављу вршило се на рачунару корисничког типа са графичком картицом (*NVIDIA GeForce RTX 3060*) и трајало је укупно око тринаест дана ефективно, тј. нешто мање од две седмице.

8

Композитни модели засновани на паралелној перплексности

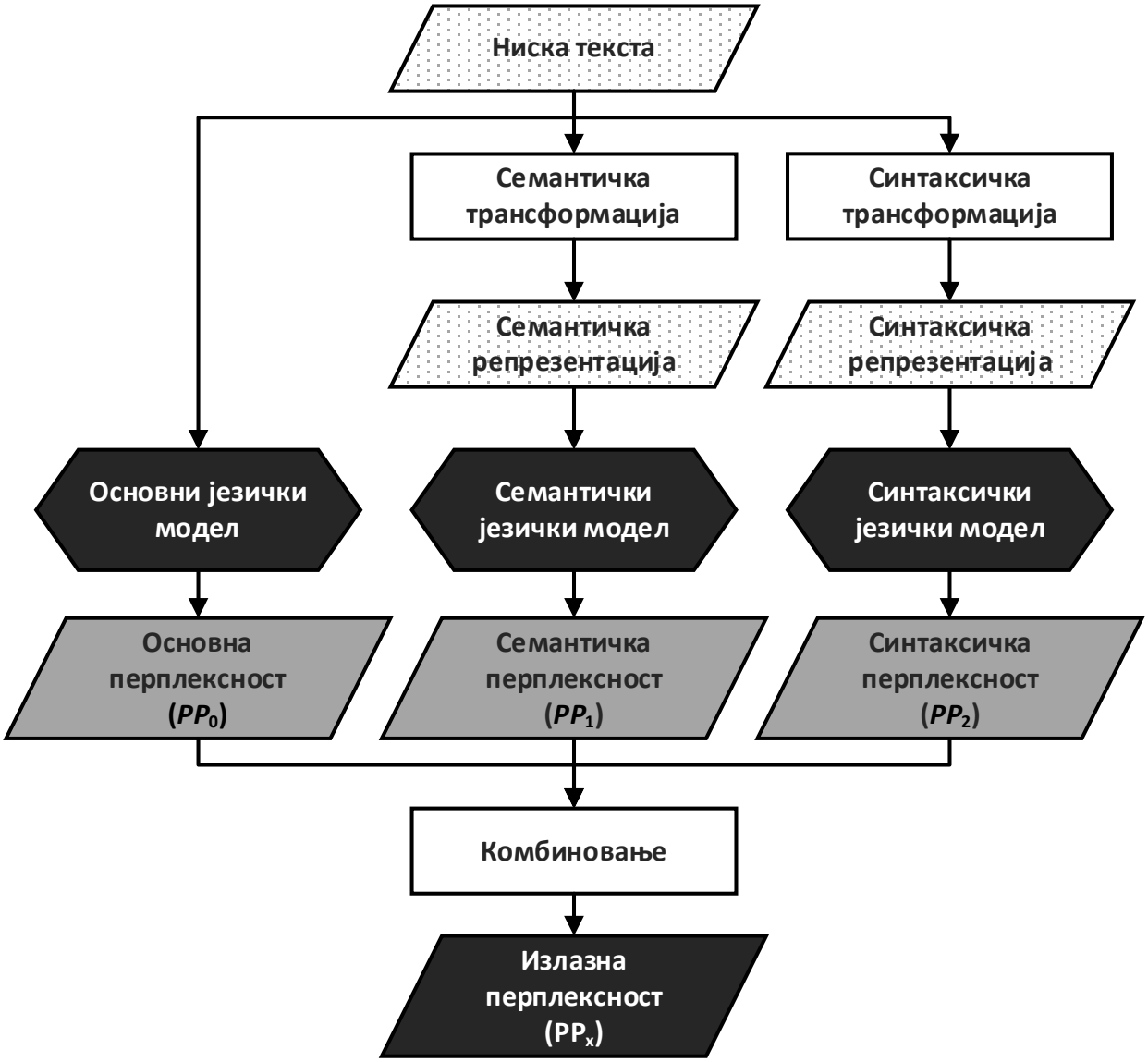
Свака композитна псеудограматика креирана за потребе овог рада биће софтверско решење које ће неки текст тј. ниску речи/токена најпре обрадити (функцијама описаним у одељцима 7.2.1 и 7.2.2) тако да се добију одговарајући улази за сваки од припремљених језичких модела (описаних у одељку 7.3), чије ће излазе (у односу на дату ниску) потом комбиновати у јединствену вредност која одражава вероватноћу задате ниске, коришћењем претходно анализираних метода комбиновања (Škorić, et al., 2022) које не повећавају димензионалност. Тако би, на пример, уколико језички модел производи скалар вероватноће, и композитни модел заснован на више таквих вредности производио такође јединствени скалар—у циљу омогућавања директног поређења између композитних и самосталних модела. Овај одељак ће, у складу са тиме, резултовати већим бројем различитих композиција које би могле бити тестиране на различитим задацима и међусобно поређене.

8.1 Перплексност у служби композитне евалуације текста

При комбиновању за потребе овог истраживања, користићемо методологију изучену у одељку 6: употребићемо језичке моделе да обрадимо текст, како би добили излазну вредност, а онда ћемо излазне вредности више модела комбиновати на различите начине: усредњавањем, множењем, минимизирањем, максимизирањем, израчунавањем векторске норме или коришћењем наслаганог класификатора. Што се тиче излаза који производе самостални модели фокусираћемо се у потпуности на меру перплексности текста (*PP*), за коју смо већ рекли да је пропорционална ентропији и израчунава се као:

$$PP = \sqrt[n]{\frac{1}{P(w_1 w_2 \dots w_n)}}$$

где је $(w_1 w_2 \dots w_N)$ ниска токена, P њихова вероватноћа, а n дужина ниске (број токена, при чему токен може бити било која подјединица текста). Са тим на уму, евалуација перплексности текста коришћењем псеудограматика реализоваће се у три корака (Илустрација 28):



Илустрација 28: Модел псеудограматике српског језика (у служби евалуације) засноване на три обучена језичка модела.

1. Свака ниска текста која је предмет евалуације биће обрађена кроз семантичку и синтаксичку трансформацију описану у претходном одељку, како би се добиле њене одговарајуће репрезентације;

2. Семантичка репрезентација улазне ниске биће прослеђена семантичком језичком моделу, синтаксичка репрезентација биће прослеђена синтаксичком, док ће основном моделу бити прослеђена неизмењена улазна ниска. Сваки од модела ће на основу добијеног улаза израчунати јединствену меру перплексности;
3. Вредности добијене у претходном кораку ће бити комбиноване тако да се добије једна, финална вредност.

8.2 Основне композиције

Као што је раније напоменуто коришћено је пет једноставних метода комбинације излазних перплексности према претходно приказаном алгоритму (Илустрација 28), како би се за сваки комад текста добило пет нових вредности перплексности $PP_x, x \in \{\text{просек}, \text{производ}, \text{минимум}, \text{максимум}, \text{внорма}\}$ које одражавају пет основних композитних модела. Сваки од њих дефинисан је на основну језичких модела које употребљава и методе комбинације свих перплексности израчунатих помоћу њих:

1. Усредњавање тј. израчунавање аритметичке средине свих израчунатих вредности перплексности:
$$PP_{\text{просек}} = \frac{1}{n}\sum_{i=0}^{n-1} PP_i = \frac{PP_0 + PP_1 + PP_2}{3}$$
2. Множење свих израчунатих перплексности:
$$PP_{\text{производ}} = \prod_{i=0}^{n-1} PP_i = PP_0 * PP_1 * PP_2$$
3. Минимизирање тј. проналажење минималне израчунате перплексности:
$$PP_{\text{минимум}} = \min_{0\le i\le 2}(PP_i) = \min(PP_0, PP_1, PP_2)$$
4. Максимизирање тј. проналажење максималне израчунате перплексности:
$$PP_{\text{максимум}} = \max_{0\le i\le 2}(PP_i) = \max(PP_0, PP_1, PP_2)$$
5. Израчунавање векторске (l^2) норме добијеног низа перплексности:
$$PP_{\text{внорма}} = \sqrt[2]{\sum_{i=0}^{n-1} {PP_i}^2} = \sqrt[2]{({PP_0}^2 + {PP_1}^2 + {PP_2}^2}$$

Приказане формуле односе се, у овом случају, на композитне моделе засноване на сва три припремљена језичка модела, што не мора нужно бити случај. Услед једноставности, за сваки приказан метод комбинације припремљене су три додатне композиције, засноване на по два језичка модела (минимални захтев), што је резултовало у укупно двадесет композитних модела (Табела 18).

Табела 18: Листа композиција језичких модела, начина њиховог компоновања и коришћених језичких модела за сваку од њих, при чему 0 означава основни, 1 семантички, а 2 синтаксички језички модел припремљен за потребе овог истраживања.

Назив композиције	**Метод компоновања**	**Коришћени модели**
к-просек $_{0+1+2}$	Усредњавање низа вредности	0, 1, 2
к-просек $_{0+1}$		0, 1
к-просек $_{0+2}$		0, 2
к-просек $_{1+2}$		1, 2
к-производ $_{0+1+2}$	Множење вредности	0, 1, 2
к-производ $_{0+1}$		0, 1
к-производ $_{0+2}$		0, 2
к-производ $_{1+2}$		1, 2
к-минимум $_{0+1+2}$	Проналажење минимума низа вредности	0, 1, 2
к-минимум $_{0+1}$		0, 1
к-минимум $_{0+2}$		0, 2
к-минимум $_{1+2}$		1, 2
к-максимум $_{0+1+2}$	Проналажење максимума низа вредности	0, 1, 2
к-максимум $_{0+1}$		0, 1
к-максимум $_{0+2}$		0, 2
к-максимум $_{1+2}$		1, 2
к-внорма $_{0+1+2}$	Израчунавање векторске (l^2) норме низа вредности	0, 1, 2
к-внорма $_{0+1}$		0, 1
к-внорма $_{0+2}$		0, 2
к-внорма $_{1+2}$		1, 2

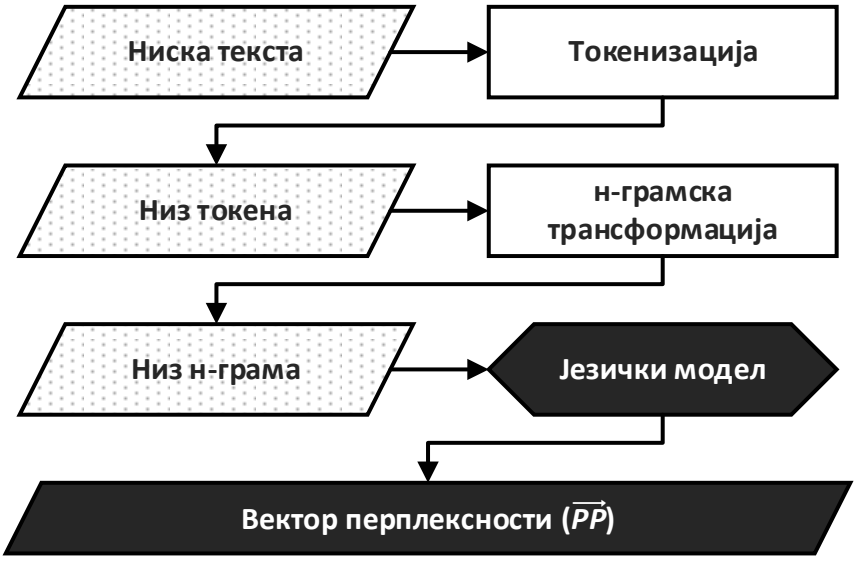
8.3 Вектори перплексности

Како се вероватноћа (и перплексност) неког текста израчунава на основу производа вероватноћа појединачних токена, може се десити да један мало вероватан токен значајно умањи вероватноћу (тј. повећа перплексност) неке иначе веома вероватне

ниске, при чему тај токен може представљати и обичну штампарску грешку. Дакле, с обзиром на то да је перплексност скаларна вредност која се односи на целу ниску, приликом њеног израчунавања губи се информација о девијацији вероватноћа појединачних токена, где релативно добар текст који има један мало вероватан токен и други текст у коме је сваки токен подједнако вероватан могу имати исту вредност перплексности, што може представљати потенцијално велики губитак информација, поготово за дуже текстове.

Са тиме на уму, као алтернатива скаларне перплексности, а за потребе овог истраживања и у циљу додатне евалуације, осмишљен је алгоритам за израчунавање вектора перплексности, својеврсних серијских репрезентација перплексности текста. Овакве репрезентације узимају у обзир претходно поменуте аспекте, те се уместо јединствене вредности израчунава релативна перплексност сваког токена текста, а те вредности се спајају у један вектор $\overrightarrow{PP}$.

Ови вектори се израчунавају у односу на целокупну ниску текста која се посматра, језички модел који одређује висину перплексности и клизећи прозор (задате) фиксне дужине *n* на следећи начин (Илустрација 29):



Илустрација 29: Израчунавање вектора перплексности за улазну ниску текста коришћењем н-грамске трансформације и предобученог језичког модела.

1. Текст се најпре подели на низ токена $w_1 w_2 \dots w_N$, где свако *w* представља по један токен, а *N* је њихов укупан број. Треба напоменути и то да се токен не мора нужно односити на токене из речника припремљеног за потребе трансформера, већ то могу бити друге речи (и подречи), фразе, или чак реченице. За потребе овог експеримента један токен ћемо поистоветити са једном речи.
2. Издваја се један по један н-грам величине *n* и прослеђује се језичком моделу који врши обраду. На пример, уколико је изабран прозор величине *n=3,* први н-грам ће се састојати од прва три токена, други од друга три токена итд. ($w_1 w_2 w_3$, $w_2 w_3 w_4$, … $w_{N-2} w_{N-1} w_N$), тако да ће укупан број н-грама бити *N – n + 1,*

а задатак језичког модела је да израчуна меру перплексности за сваки од њих појединачно.
3. За сваки токен се рачуна упросечена перплексност свих н-грама којима припада, а низ перплексности свих токена представља коначну серију.

На овај начин, и коришћењем раније обрађиване ниске (Пример 7), основног језичког модела и прозора величине *n=5* долазимо до следећих вредности (Табела 19):

Табела 19: Вредности перплексности сваког појединачног 5-грама токена познате изреке у односу на основни језички модел обучен за потребе овог истраживања.

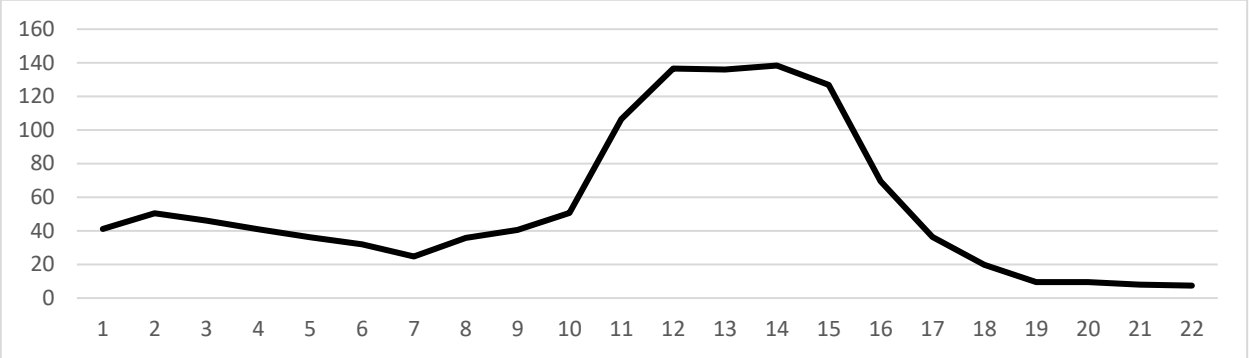
н-грам токена који се обрађује (5-грам)	**израчуната перплексност**
Ако се задесиш у Риму	41.1823
се задесиш у Риму,	59.6357
задесиш у Риму, понашај	37.6951
у Риму, понашај се	25.5811
Риму, понашај се као	17.2702
, понашај се као Римљанин	19.7407
понашај се као Римљанин.	23.2376
се као Римљанин. Ако	92.7277
као Римљанин. Ако се	49.3802
Римљанин. Ако се задесиш	67.7575
. Ако се задесиш другде	299.4118
Ако се задесиш другде,	173.6724
се задесиш другде, понашај	89.9541
задесиш другде, понашај се	61.0136
другде, понашај се према	9.8791
, понашај се према тамошњим	12.5875
понашај се према тамошњим обичајима	8.4260
се према тамошњим обичајима.	7.4211

Када се на основу добијених вредности за сваки од 5-грама израчунају перплексности за сваки токен (усредњавањем перплексности свих 5-грама у којима се налази), добијају се следеће вредности (Табела 20):

Табела 20: Вредности перплексности сваког појединачног токена познате изреке (добијене експерименталном методом) у односу на основни језички модел обучен за потребе овог истраживања.

редни број	**токен**	**израчуната перплексност**
1	*Ако*	41.1824
2	*се*	50.4091
3	*задесиш*	46.1711
4	*у*	41.0236
5	*Риму*	36.2729
6	*,*	31.9846
7	*понашај*	24.7050
8	*се*	35.7115
9	*као*	40.4713
10	*Римљанин*	50.5688
11	*.*	106.5030
12	*Ако*	136.5899
13	*се*	136.0352
14	*задесиш*	138.3619
15	*другде*	126.7862
16	*,*	69.4214
17	*понашај*	36.3721
18	*се*	19.8655
19	*према*	9.5785
20	*тамошњим*	9.4783
21	*обичајима*	7.9236
22	*.*	7.4211

Приказане вредности се потом могу ефективно визуелно приказати пресликавањем вредности перплексности на *y* осу и токена (њихових редних бројева) на *x* осу. На тај начин добијамо линијски графикон која нам директно показује који делови текста одражавају највишу, а који најнижу перплексност, као и њену девијацију (Илустрација 30).



Илустрација 30: Вектор перплексности приказан у виду линијског графа, где *x* оса одражава ток текста (реченице), а *y* оса одражава меру перплексности.

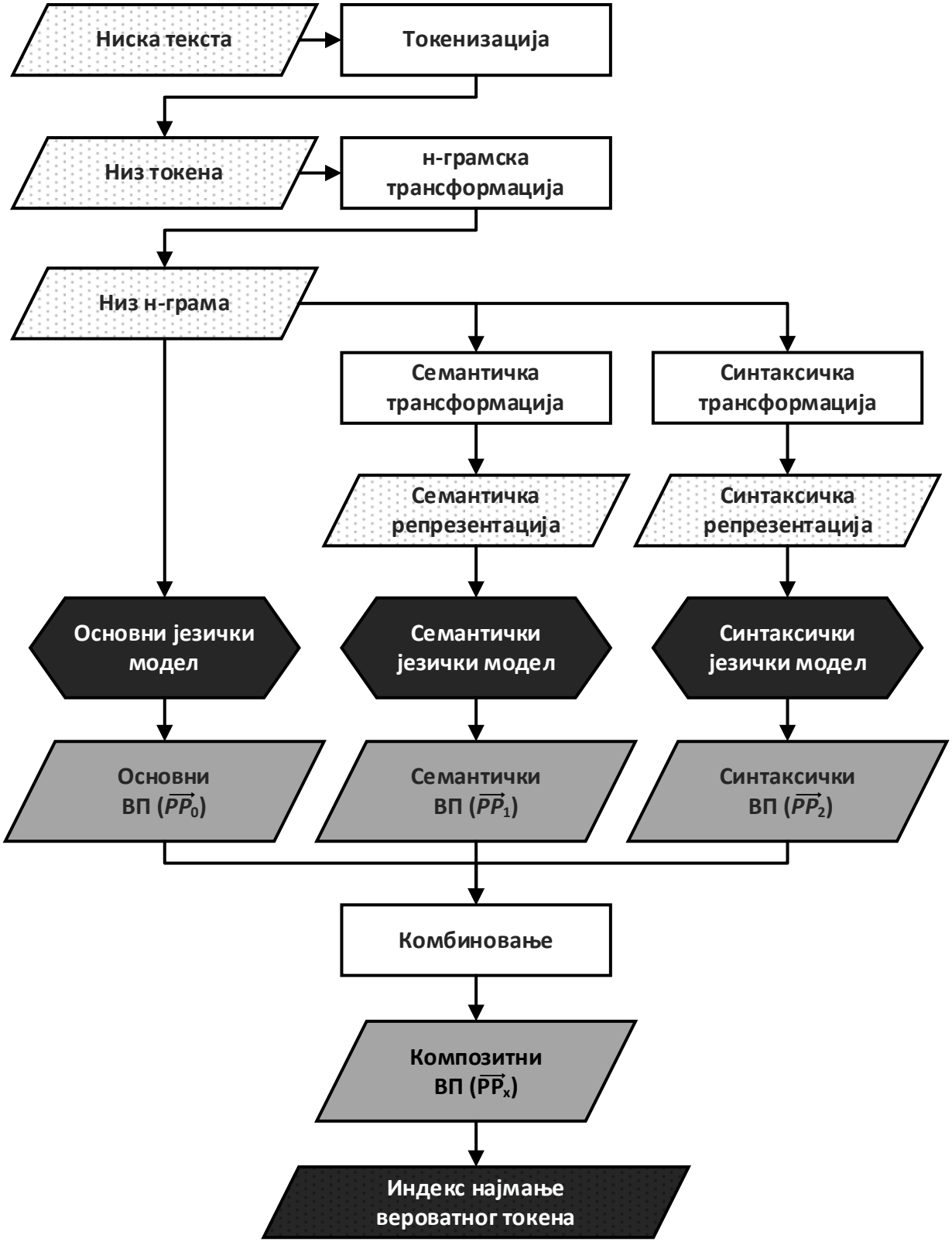
Осим потенцијално бољег моделирања перплексности неког текста, овај приступ, дакле, омогућава и директну детекцију речи или делова текста са највећом мером перплексности, који могу представљати потенцијално исправљиве грешке, а комбинацијом оваквих вектора, добијених на основу више језичких модела, се резултати могу и додатно побољшати.

8.4 Композиције засноване на векторима перплексности

8.4.1 Вектори перплексности у служби детекције грешака у тексту

Као што је напоменуто у одељку 8.3, вектори перплексности (ВП) се могу потенцијално користити и за проналажење грешака у тексту, лоцирањем делова текста који имају највећу релативну перплексност. Композиције засноване на претходно израчунатим ВП су за те потребе добијене серијском применом алгоритма за израчунавање ВП (Илустрација 29) и алгоритма за комбиновање перплексности у служби евалуације (Илустрација 28). Процес добијања јединствене ВП се, у складу са тиме, одвија на следећи начин (Илустрација 31):

1. Улазни текст се најпре подели на низ токена $w_1 w_2 \dots w_N$, где свако w представља по један токен, а N представља њихов укупан број. Као што је већ напоменуто, токени се не морају нужно односити на токене из речника припремљеног за потребе трансформера, већ то могу бити друге речи, фразе, реченице итд.;
2. Креира се низ свих н-грама добијених токена величине n, за неко задато n. Уколико је изабран прозор величине n=3, први н-грам ће се састојати од прва три токена, други од друга три токена итд. ($w_1 w_2 w_3$, $w_2 w_3 w_4$, $\dots w_{N-2} w_{N-1} w_N$), тако да ће укупан број н-грама бити $N - n + 1$;
3. Сваки од добијених н-грама токена обрађује се кроз претходно описане семантичке и синтаксичке трансформације (одељци 7.2.1 и 7.2.2), како би се добиле одговарајуће репрезентације;
4. Семантичке репрезентације н-грама прослеђују се семантичком језичком моделу, синтаксичке репрезентације синтаксичком, док ће основном моделу бити прослеђен неизмењен низ н-грама токена. Сваки од модела ће за сваки добијени улаз израчунати јединствену меру перплексности;



Илустрација 31: Комбинација вектора перплексности (ВП) израчунатих помоћу различитих језичких модела, њихово комбиновање и композитно-заснована детекција индекса најмање вероватног токена, тј. токена који одаје највећу релативну перплексност у задатом тексту.

5. За сваки језички модел се на основу добијених перплексности н-грама из претходног корака израчунава ВП, тако што се за сваки токен из првог корака рачуна упросечена додељена перплексност (додељена од стране специфичног језичког модела) свих н-грама којима припада. Добијени низ представља коначан ВП за сваки језички модел.
6. ВП добијени у претходном кораку се комбинују у јединствени вектор коришћењем установљених метода комбинације, при чему се комбинују перплексности појединачних токена, и тако добијене вредности чине коначне, композитно-израчунате ВП.

Овакви модели се могу користити, између осталог, у сврхе:

- Детекције погрешне речи у тексту;
- Детекцију места у тексту на коме реч недостаје;
- Детекцију места у тексту на коме је грешком уметнута реч.

За сваки од поменутих примера биће припремљени одговарајући скупови за евалуацију, а биће евалуирано укупно двадесет композитних модела, по четири (за четири различите комбинације модела, 0+1+2, 0+1, 0+2 и 1+2) за сваки од пет начина комбиновања (усредњавање низа вредности, множење вредности, екстракција минимума низа вредности, екстракција максимума низа вредности и израчунавање векторске (l^2) норме низа вредности), названи по принципу:

$$\text{д-}[\textit{метод компоновања}]_{[\textit{коришћени модели}]}$$

тј.

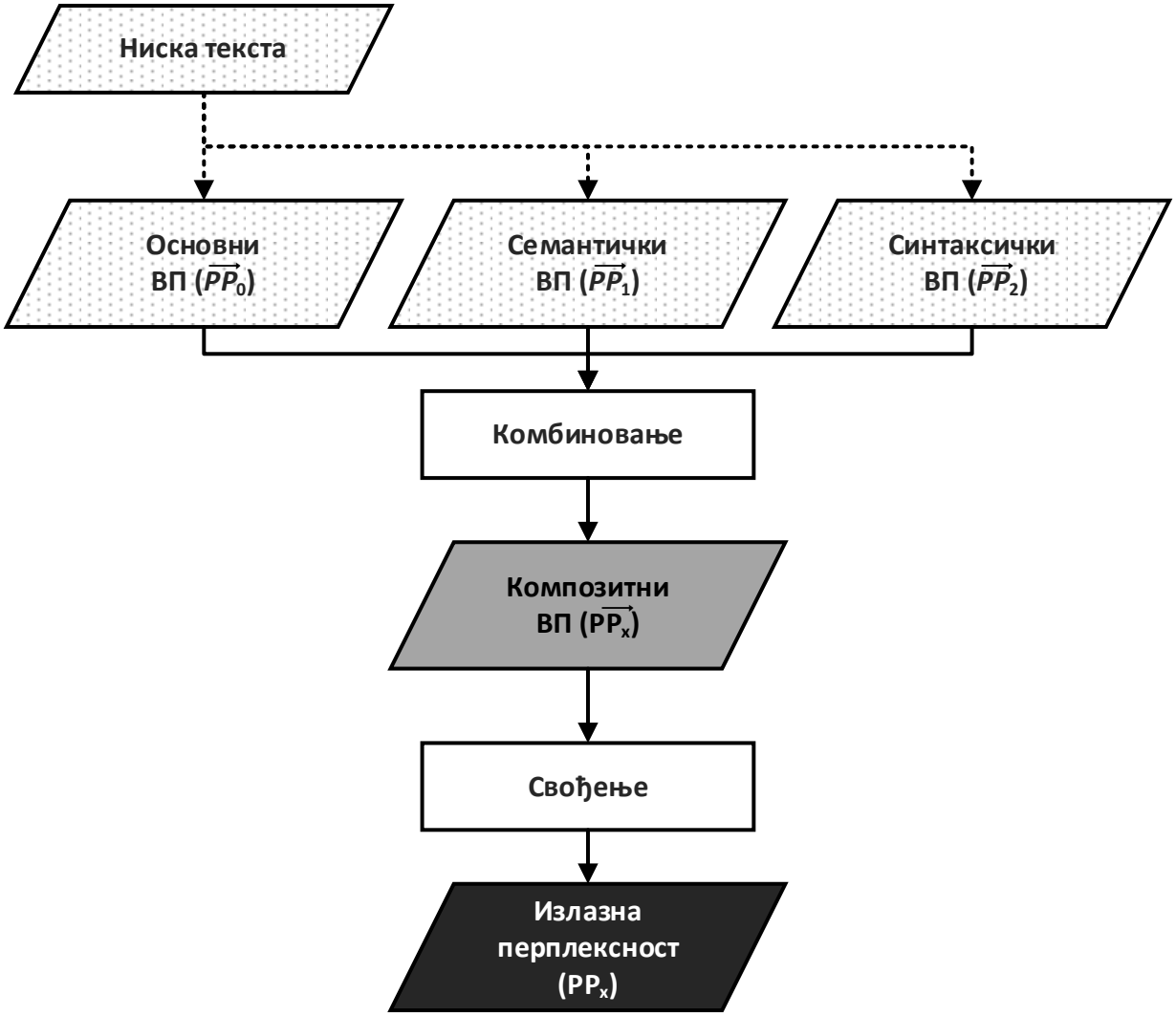
$$\text{д} - \begin{bmatrix} \text{просек} \\ \text{производ} \\ \text{минимум} \\ \text{максимум} \\ \text{внорма} \end{bmatrix} \begin{bmatrix} 0+1+2 \\ 0+1 \\ 0+2 \\ 1+2 \end{bmatrix}$$

и где би модел у служби детекције најмање вероватног токена, заснован на сва три обучена језичка модела и комбиновањем адекватних ВП усредњавањем био обележен као:

$$\text{д-просек}_{\ 0+1+2}$$

8.4.2 Сажети вектори у служби евалуације перплексности

У циљу што комплетније компаративне анализе развијен је додатни приступ композитног израчунавања перплексности (Илустрација 32) који се ослања на претходно израчунате ВП, претходно описаном комбиновању ВП (Илустрација 31) и њиховом *свођењу* тј. израчунавању јединствене скаларне вредности, поново помоћу неког од описаних начина комбиновања више вредности у једну.



Илустрација 32: Израчунавање јединствене вредности перплексности на основу више ВП, најпре њиховим комбиновањем у један, те свођењем на скаларну вредност, при чему је испрекиданом линијом представљен сажет процес добијања адекватних ВП за неку задату ниску текста.

На пример, рецимо да имамо три ВП, израчуната коришћењем три различита обучена модела

$$\overrightarrow{PP_0} = (1,2,3,1)$$

$$\overrightarrow{PP_1} = (2,3,4,3)$$

$$\overrightarrow{PP_2} = (2,5,2,6)$$

те да за композицију користимо два различита начина: усредњавање и множење. Помоћу приказаног алгоритма добили бисмо најпре (коришћењем усредњавања према позицији) један јединствени вектор:

$$\overrightarrow{PP_x} = (\tfrac{1}{3}\sum(1,2,2)\,,\tfrac{1}{3}\sum(2,3,5)\,,\tfrac{1}{3}\sum(3,4,2)\,,\tfrac{1}{3}\sum(1,3,6))$$

$$\overrightarrow{PP_x} = (\sim 1.67, \sim 3.33, 3, \sim 3.33)$$

који одговара аритметичкој средини најпре свих првих, потом свих других, свих трећих и свих четвртих вредности координата вектора, а од којег би даље добили производ у вредности:

$$PP_x = \prod(\sim 1.67, \sim 3.33, 3, \sim 3.33) = \sim 55.5$$

Ако би обрнули редослед операција добили најпре вектор производа према позицијама:

$$\overrightarrow{PP_x} = (\prod(1,2,2)\,, \prod(2,3,4)\,, \prod(3,4,2)\,, \prod(1,3,6))$$

$$\overrightarrow{PP_x} = (4, 30, 24, 18)$$

коју би потом усреднили у вредност

$$PP_x = \frac{1}{4}\sum(4, 30, 24, 18) = 19$$

На овај начин, коришћењем пет различитих начина комбиновања и сажимања добија се укупно 25 композиција које дају потенцијално различите резултате. Када се тих 25 композиција примени на четири различите комбинације модела (0+1+2, 0+1, 0+2 и 1+2) добија се укупно 100 нових композитних модела погодних за тестирање, названих на следећи начин:

к-[*први метод*]-[*други метод*]$_{[\text{коришћени модели}]}$

тј.

$$\text{д} - \begin{bmatrix} \text{просек} \\ \text{производ} \\ \text{минимум} \\ \text{максимум} \\ \text{внорма} \end{bmatrix} - \begin{bmatrix} \text{просек} \\ \text{производ} \\ \text{минимум} \\ \text{максимум} \\ \text{внорма} \end{bmatrix} \begin{bmatrix} 0+1+2 \\ 0+1 \\ 0+2 \\ 1+2 \end{bmatrix}$$

Дакле, композитни модел који користи векторе перплексности, тако што их најпре комбинује па сажима, користи усредњавање (*просек*) за комбиновање и множење вредности (*производ*) за сажимање, и користи сва три предобучена трансформера назива се:

к--просек-производ $_{0+1+2}$

8.5 Композиције модела у служби генерисања текста

Иако можемо директно сравњивати вероватноће текста израчунате помоћу различитих модела када је у питању евалуација, то не важи и за задатак генерисања текста. Иако сви модели користе исти *huggingface* токенизатор и речник токена, обучени су на различитим репрезентацијама текста, па тако и генеришу различите репрезентације, као што је претходно илустровано (поглавље 7.3: Пример 10, Пример 11 и Пример 13).

Уколико би се комбиновале вероватноће токена у циљу генерисања наставка неког контекста, добио би се текст који је комбинација тих репрезентација, што није оптимално решење. На примеру генерисања текста коришћењем усредњене вероватноће израчунате помоћу сва три обучена модела (Пример 14) приметно је да се највише испољава синтаксичка репрезентација.

статус Рёüй Рìýú ßёüû од сеÞòìüÿû населити öòýû ßìÿú аеßòăû adêëþû, öñÞёÿёù Рíÿñù, Òёß је на Рсú у Рú,

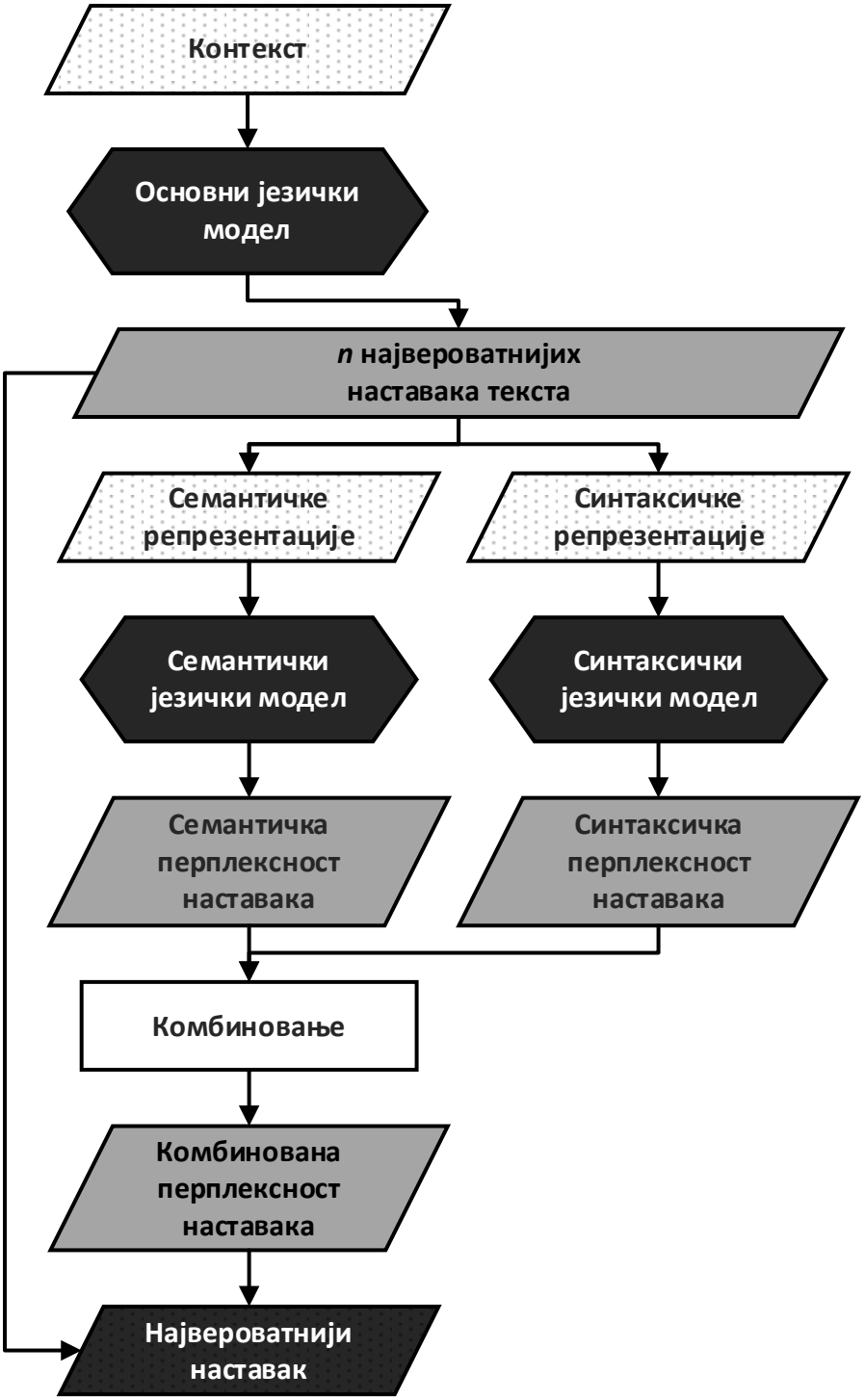
испунити Рёÿú Рìăú ÒёÞ је да је ßёüú ßìýú па Рíÿù Òìê da bude öòßёýû ßíþú, али да је öñÞёсû öёòñêёс Рú êё1q ßú у

бер Рёüù Рìăú Рíÿú Èôì ßёăú аеÞÿû ßìýú ßíþú Òёß је Рú öòßёýû öñÞёÿù, ßăú adêëþû, ÞаеÞă

Пример 14: Пример три исечка текста који су генерисани на основу композиције матрица вероватноће токена израчунате коришћењем језичких модела обучених на различитим репрезентацијама текста.

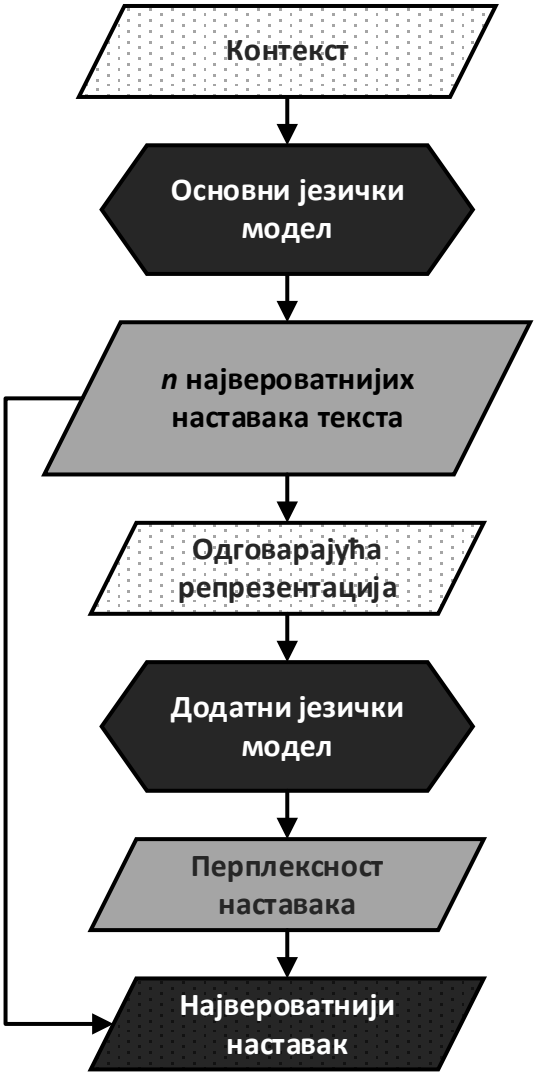
Како алтернатива овом приступу осмишљен је алгоритам за генерисање заснован на правилима композиције модела, који генерише читљив текст и који се извршава у три корака (Илустрација 33):

1. На основу задатог контекста (или насумичним одабиром првог токена) и неке вредности *n*, основни језички модел генерише листу од *n* могућих наставака текста;
2. Кандидати се најпре трансформишу у одговарајуће репрезентације и евалуирају коришћењем семантичког и синтаксичког модела, а израчунате перплексности се комбинују, неким од претходно установљених начина комбиновања;
3. За коначан излаз узима се наставак текста који има најмању комбиновану перплексност.



Илустрација 33: Генерисање наставка текста засновано на генерисању *n* кандидата и њиховом композитном евалуацијом како би се пронашао најподобнији кандидат.

Такође, наставак текста може бити генерисан и коришћењем само једног додатног модела за евалуацију. Том приликом се наставак бира искључиво засновано на перплексности коју је израчунао додатни модел и не користи се ниједан од метода комбинације (Илустрација 34).



Илустрација 34: Генерисање наставка текста помоћу перплексности израчунатих коришћењем само једног додатног модела за евалуацију.

Коришћењем ова два приступа добија се укупно седам јединствених композитних модела за генерисање, од којих је пет засновано на пет комбинација семантичких и синтаксичких перплексности кандидата и преостала два заснована на коришћењу по једног модела за евалуацију (Табела 21).

Табела 21: Листа композиција језичких модела у служби генерисања, начина њиховог компоновања и коришћених језичких модела за сваку од њих, при чему 0 означава основни, 1 семантички, а 2 синтаксички језички модел припремљен за потребе овог истраживања.

Назив композиције	**Метод компоновања**	**Коришћени модели**
г-просек $_{0+1+2}$	Усредњавање низа вредности	0, 1, 2
г-производ $_{0+1+2}$	Множење вредности	0, 1, 2
г-минимум $_{0+1+2}$	Екстракција минимума низа вредности	0, 1, 2

г-максимум $_{0+1+2}$	Екстракција максимума низа вредности	0, 1, 2
г-внорма $_{0+1+2}$	Израчунавање векторске (l^2) норме низа вредности	0, 1, 2
г $_{0+1}$	без компоновања	0, 1
г $_{0+2}$	без компоновања	0, 2

Примери наставака текста генерисаних од стране ових седам композитних модела за задати леви контекст

Ако се задесиш у Риму, понашај се као Римљанин. Ако се задесиш другде,

приказани су испод (Табела 22).

Табела 22: Примери по три генерисана наставка текста за сваки од седам различитих композитних модела за генерисање, а за задати леви контекст у виду почетка познате пословице *„Ако се задесиш у Риму, понашај се као Римљанин. Ако се задесиш другде,“* .

Назив композиције	**Пример генерисаног наставка текста**
г-просек $_{0+1+2}$	*онда си проглашен за професионалног војника. Ако прођеш кроз процес производ* *онда си пријатељ Рима. Ако си преварант, пратиш људе и прислушкујеш* *онда се употребиш као војник. Ако употребиш војнички израз,*
г-производ $_{0+1+2}$	*онда си пријатељ Рима. Ако си припадник италијанске војске, присталица је* *онда си присиљен да се придружиш пријатељима. Ако је прогонство* *онда си припадник исте етничке групе. Ако прихватите да је припадала истом народу,*
г-минимум $_{0+1+2}$	*онда си присиљен да се придружиш пријатељима. Ако не прихватиш* *онда си принуђен да се вратиш у Италију. Ако престанеш да се бавиш политиком,* *онда си присталица проклетства. Ако си противник пророка, прогонитеља,*

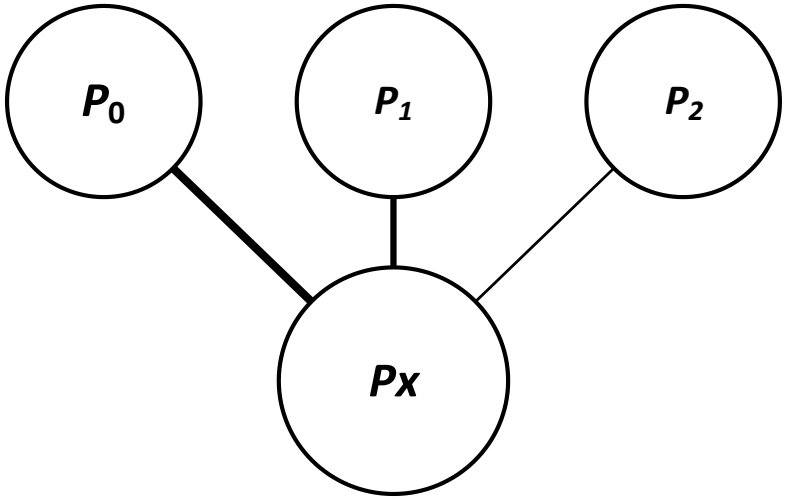
г-максимум $_{0+1+2}$	*онда си пријатељ престолонаследника.* *Ако је у Рим стигао принц од Велса,* *онда си пријатељ Рима. Ако си припадник италијанске војске, прво си војник* *онда си принуђен да се вратиш у Италију.* *Ако си прекршио правила, престани*
г-внорма $_{0+1+2}$	*онда си пријатељ Рима. Ако не прихватиш да будеш присталица Рима, престани* *онда си принуђен да останеш у граду.* *Ако не останете у Римском царству, престани* *онда си прво присталица првог принципа.* *Ако си први пријатељ принца,*
г $_{0+1}$	*онда си пријатељ Рима. Ако не прихватиш да будеш присталица Рима, престани* *онда си принуђен да се вратиш у Италију.* *Ако си прекршио правила, престани* *онда си пријатељ Рима. Ако прихватиш да будеш присталица Рима, престани да*
г $_{0+2}$	*онда си присиљен да се вратиш у Италију.* *Ако си преварен, престани да се бави* *онда си принуђен да се вратиш у Италију.* *Ако си прекршио правила, престани* *онда си присиљен да се вратиш у Италију.* *Ако не прихватиш принцип слобо*

8.6 Композиције засноване на вештачким неуронским мрежама

Поред композиција заснованих на правилима описаним у поглављима од 8.1 до 8.5, предвиђено је и креирање и тестирање композиција заснованих на наслаганим класификаторима (вештачким неуронским мрежама) креираних на начин претходно приказан у поглављима 5.3 и 6.3.

У случају коришћења наслаганог перцептрона као метода композиције јединствених перплексности, вредности добијене од стране три трансформера (PP_0, PP_1 и PP_2) су

конвертоване у вероватноће (P_0, P_1 и P_2) и коришћене за обучавање и евалуацију перцептрона који одају јединствену меру вероватноће задате ниске (Илустрација 35).



Илустрација 35: Наслагани перцептрон обучен да на основу три вредности вероватноћа израчунатих од стране различитих модела одреди јединствену меру вероватноће.

Перцептрони за потребе овог истраживања обучени су на задатку унарне класификације (моделовања језика), тако да директно врше функцију њихове псеудограматике где излазна вредност одговара вероватноћи да унета ниска текста припада језику који се моделује. Наслагани класификатори су обучавани да разлуче између експертских и машинских превода књижевних дела (о којима ће бити више речи у наредном одељку). Њима су тако приликом обучавања прослеђиване реченице које припадају или не припадају предвиђеном мини-језику, где су прве обележене са 1, друге са 0, а карактеристике које су коришћене за обучавање су биле вероватноће додељене тим реченицама од стране предобучених језичких модела (P_0, P_1 и P_2, Илустрација 35).

Као што је претходно описано и у одељку 6.4.2, за обучавања је поново коришћен оптимизатор *ADAM* (Kingma & Ba, 2014) са почетном стопом учења (*initial learning rate*) од 0.01, величином серија (*batch size*) фиксираном на 64, а обучавања су трајала по 100 епоха. За потребе овог експеримента обучено је и евалуирано путем петоструке унакрсне валидације пет класификатора, дакле, кроз пет итерација перцептрони су обучавани на различитих четири петина парова реченица, а њихов учинак је евалуиран на одговарајућој петој петини. Добијене композиције обележене су у складу са бројем итерације:

$$\text{к} - \text{перцептрон} - \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

Када је наслагани класификатор обучен, алгоритам за композитно израчунавање вероватноће да нека реченица припада предвиђеном језику се реализује на следећи начин (Илустрација 36):

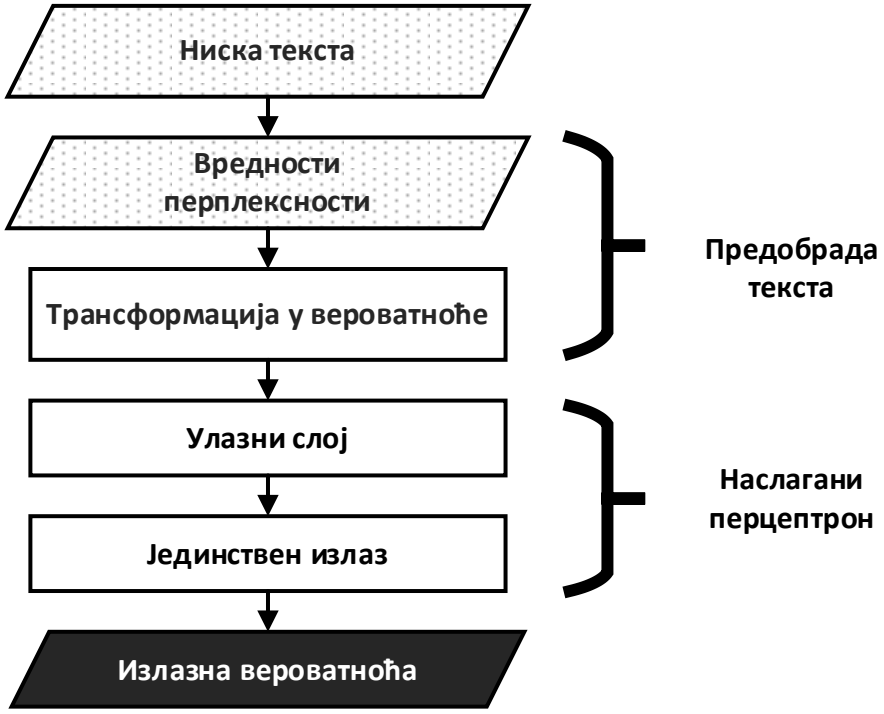
- Предобрада. У овом кораку се најпре одвија израчунавање перплексности (основне, семантичке и синтаксичке тј. PP_0, PP_1 и PP_2) за задату ниску текста коришћењем претходно утврђеног алгоритма (Илустрација 28) и одговарајућих предобучених језичких модела: основног, семантичког и синтаксичког. Добијене перплексности се потом трансформишу у вероватноће узимањем реципрочне верзије сваког појединачног броја у серији

$$x = \frac{1}{x}$$

  како би добиле вредности између 0 и 1

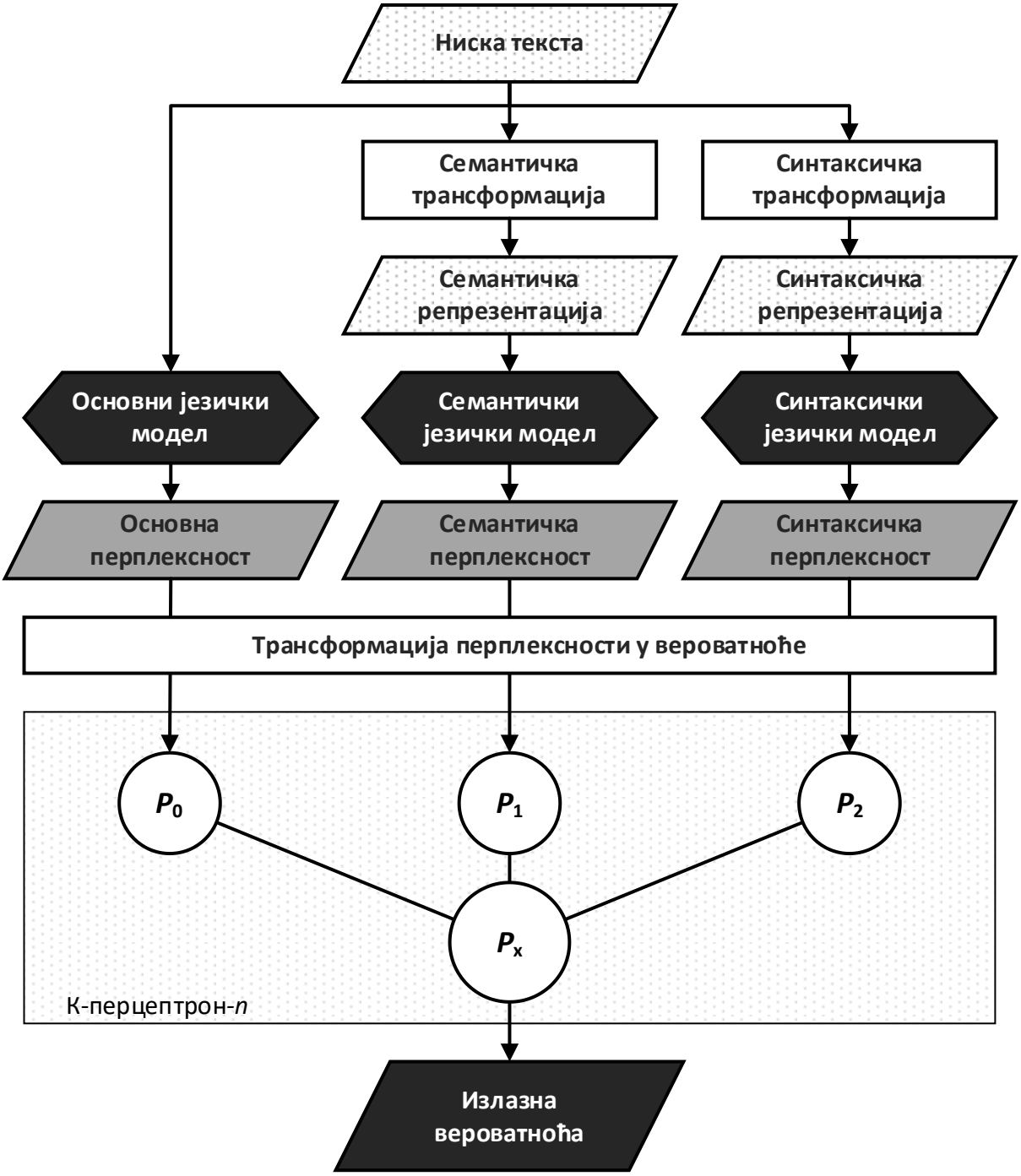
$$(\forall x \in \{PP_0, PP_1, PP_2\})(0 \leq x \leq 1).$$

- Обрада наслаганим перцептроном. Реализација се завршава коришћењем карактеристика произведених у претходном кораку као улаза за претходно обучени перцептрон, који на основу три улазне производи једну коначну вероватноћу.



Илустрација 36: Израчунавање вероватноће текста коришћењем јединствених вредности перплексности као улаза за перцептрон са једним излазним чвором.

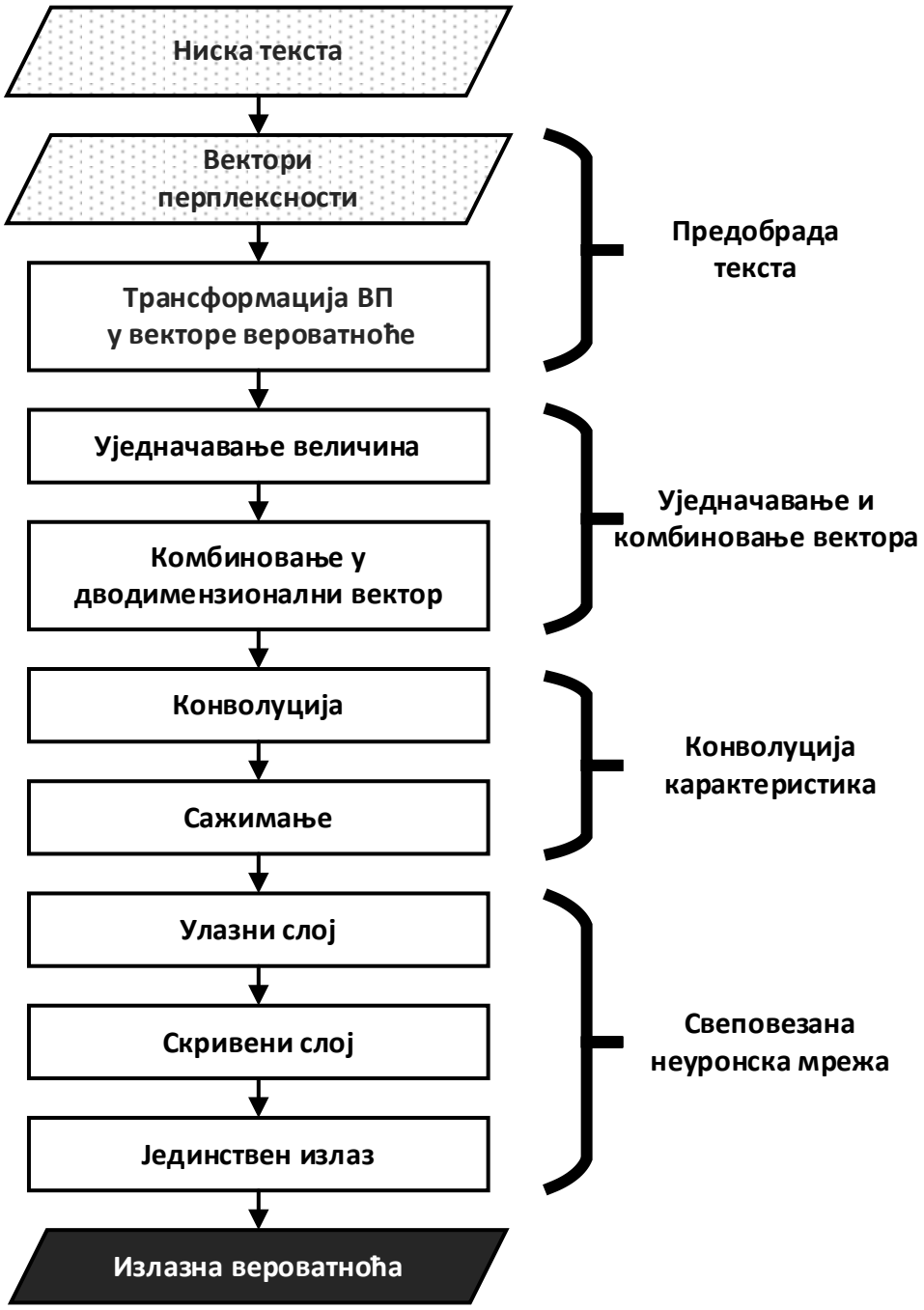
Комплетан модел псеудограматике засноване на комбиновању језичких модела коришћењем наслаганог перцептрона приказан је у наставку (Илустрација 37).



Илустрација 37: Модел псеудограматике једног мини-језика (у служби евалуације) засноване на три обучена језичка модела и наслаганом перцептрону.

У случају да приказана архитектура не буде испунила очекивања, или уколико се кроз евалуацију на неком од осмишљених задатака покаже да она нема довољан капацитет за адекватно моделовања језика, припремљена је и алтернатива, заснована на раније поменутим векторским репрезентацијама перплексности (одељак 8.3) док се као наслагани класификатори користе конволуционе неуронске мреже (*Convolutional neural network*) (O'Shea & Nash, 2015), оптимално решење у обради вишедимензионалних вектора, које се превасходно користе за обраду слика, али све

више налазе примену и у обради векторизованог текста на природном језику. Процес евалуације се одвија у четири главна корака (Илустрација 38):



Илустрација 38: Израчунавање вероватноће текста помоћу вектора вероватноћа, њихових конволуција у карактеристике и ВНМ са једним скривеним слојем и једним излазним чвором.

1. Предобрада. У овом кораку се најпре одвија израчунавање различитих ВП (основни ВП, семантички ВП и синтаксички ВП тј. $\overrightarrow{PP_0}$, $\overrightarrow{PP_1}$ и $\overrightarrow{PP_2}$) за задату ниску текста коришћењем претходно утврђеног алгоритма (Илустрација 29) и одговарајућих предобучених језичких модела: основног, семантичког и

синтаксичког. Добијени ВП се потом трансформишу у векторе вероватноће узимањем реципрочне верзије сваке појединачне координате

$$x = \frac{1}{x}$$

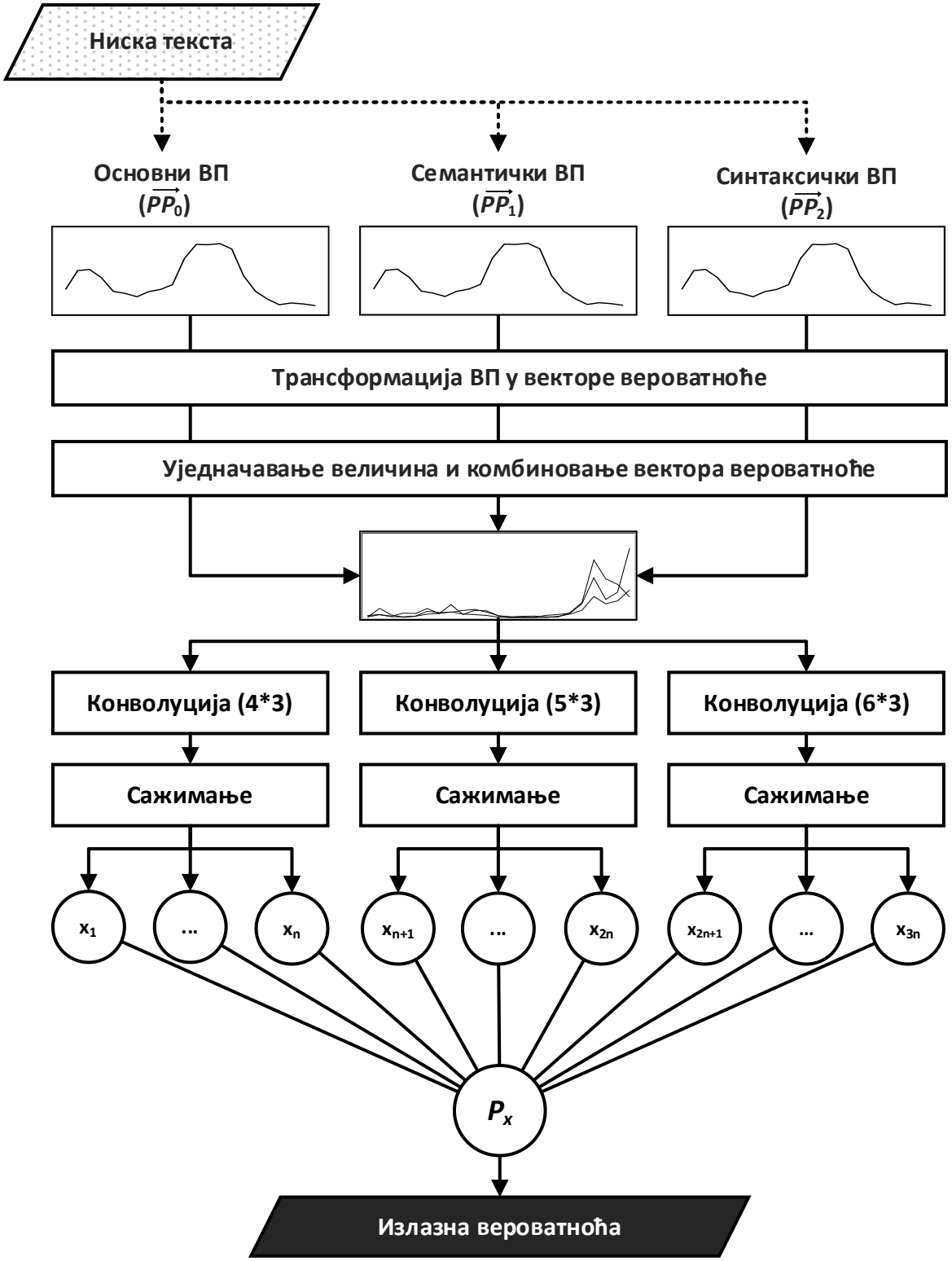
како би добили вектори са вредностима координата између 0 и 1.

2. Уједначавање и комбиновање. Како би вектори свих реченица били једнаких дужина, одређена је јединствена фиксна дужина вектора, 64. Сви вектори дужи од 64 (израчунати на основу реченица дужих од 64 речи) скраћени су на прве 64 вредности, док су вектори краћи од 64 допуњавани нултим вредностима (*zero padding*). Сва три произведена вектора дужине 64 су потом комбинована у јединствени дводимензионални векторе величине 64*3, који је коришћен у даљој обради.
3. Конволуција карактеристика. У овом кораку се над добијеним дводимензионалним вектором вероватноћа три пута примењују конволуција (*convolution*) коришћењем језгара (*kernel*) величина 4, 5 и 6 (4*3, 5*3 и 6*3) и потом сажимање (*pooling*) како би се припремио скуп карактеристика фиксне дужине n који ће бити коришћен као улаз у стандардну вештачку неуронску мрежу. Приликом сваке од конволуција користи се ход 2 (*stride*=2), док је број излазних параметара која се добијају био n=32. Овај корак тако резултује у укупно 96 (три пута по 32) параметара добијених обрадом дводимензионалног вектора и који ће се користити у даљој обради.
4. Обрада свеповезаним неуронским слојем (*fully-connected layer*). Реализација се завршава коришћењем карактеристика из претходног корака (базираних на сва три ВП) као улаза за последњи, свеповезани слој неуронске мреже који као излаз има вероватноћу да задата ниска припада језику над којим је обучен овај модел — P_x, као и код примера из претходне секције (Илустрација 37). Такође, као и у претходном случају, модел се обучава на задатку моделовања мини-језика, при чему је жељени излаз 1, уколико је у питању адекватна реченица и 0, у супротном.

Као и код наслаганих перцептрона обучавање се врши коришћењем оптимизатора *ADAM* (Kingma & Ba, 2014) са почетном стопом учења (*initial learning rate*) од 0.01, величином серија (*batch size*) фиксираном на 64. Ипак, услед комплексности мреже (и неопходности веће количине података) број епоха није био ограничен на 100, већ је обучавање заустављано када дође до дивергенције у ентропији скупова за обучавање и валидацију (*training loss* и *validation loss*), тј. када први критеријум вредност која је за 20 или више посто нижа од вредности другог критеријума. Произведено је укупно пет оваквих композиција, обележених према итерацији унутар унакрсне провере:

$$\text{к} - \text{конволуција} - \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

Коначан модел приказан је у наставку (Илустрација 39).



Илустрација 39: Модел псеудограматике једног мини-језика (у служби евалуације) засноване на три обучена језичка модела, израчунавању вектора перплексности (вероватноћа) и конволуционим неуронским мрежама, где је n=32.

9
Евалуација

Тестирање припремљених језичких модела и композитних псеудограматика засновано је на специјалним скуповима за тестирање креираним на основу паралелизованих корпуса, морфолошких речника српског језика и различитих аутоматских процедура, специјално за ово истраживање. Креирано је укупно осам скупова (на српском језику), предвиђених за тестирање припремљених модела (у циљу евалуације) на три оквирна задатка:

1. Детекција семантичких и синтаксичких неправилности
2. Детекција уклоњене, уметнуте и замењене речи
3. Моделовање мини-језика тј. детекција машинских превода реченица

Списак скупова података који су креирани за потребе истраживања као и задаци евалуације за које се скупови примењују приказани су испод (Табела 23).

Табела 23: Припремљени скупови за евалуацију и задаци за које се користе.

	Скуп за евалуацију	**Задатак за евалуацију**
1	Листа квалитетних реченица на српском језику (експертских превода) које ће се користити као златни стандард	Израчунавање основних перплексности
2	Листе окрњених реченица, где је насумично уклоњена једна реч	Детекција уклоњене, уметнуте и замењене речи
3	Листе проширених реченица, где је насумично додата једна реч из речника	

4	Листе реченица где је једна реч насумично замењена одговарајућом алтернативом	Детекција уклоњене, уметнуте и замењене речи
5	Листа реченица које нису синтаксички исправне, добијених лематизацијом	Детекција семантичких и синтаксичких неправилности
6	Листа реченица које нису синтаксички исправне, добијених насумичним мешањем речи	
7	Листа реченица које су исправне синтаксички, али не и семантички, добијена насумичном заменом речи	
8	Листа реченица која је пандан првој листи, а добијена је машинским превођењем те исте реченице са страног језика	Моделовање мини-језика / детекција машинских превода

Резултати ће бити израчунати за сваки модел и комбинацију понаособ и међусобно упоређени, у циљу изрицању коначног суда о томе да ли комбиновање различитих (и којих) језичких модела доводи до побољшања резултата на задатку моделовања српског језика (ИП4), као и који метод комбинације је најбољи (ИП5).

9.1 Припрема скупова за евалуацију

Скупови података за евалуацију креираних модела засновани су на паралелизованим корпусима књижевних текстова (књижевна дела изворно написана на неком од најзаступљенијих европских језика и њихових експертских превода на српски језик), који нису коришћени за обучавање језичких модела описаних у секцији 7.3, како би се избегао проблем пристрасности (*bias*) приликом евалуације.

Први ресурс који је коришћен је исечак паралелног српско-немачког корпуса, *СрпНемКор* (Andonovski, et al., 2019), при чему су коришћени само текстови романа изворно писаних на немачком језику. Други ресурс који је коришћен је паралелизовани превод трећег дела серијала *Напуљске приче* (Perišić, et al., 2022), објављен у оквиру паралелног српско-италијанског корпуса креираног за потребе пројекта *It-Sr-Ner*, у оквиру организације *CLARIN* (Krauwer & Hinrichs, 2014). Коришћено је укупно седам паралелизованих романа (Табела 24).

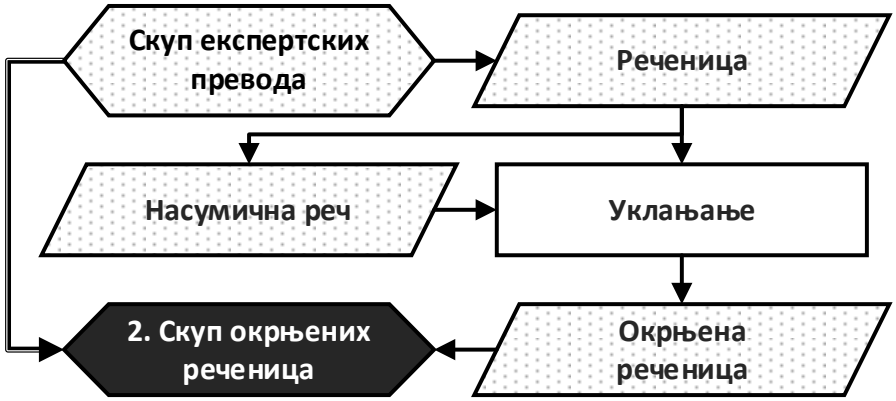
Табела 24: Паралелизовани романи коришћени за израду скупова за евалуацију, њихов аутор, наслов, изворни језик и број реченица који је ушао у корпус.

	Аутор / Преводилац	Наслов	Изворни језик	бр. реч.
1	Томас Бернхард / Бојана Денић	*Моје награде*	немачки	1009
2	Елфриде Јелинек / Тијана Тропин	*Пијанисткиња*		6679
3	Мило Дор / Томислав Бекић	*Беч, јули 1999*		1249
4	Гинтер Грас / Александра Гојков Рајић	*Ходом рака*		2868
5	Гинтер де Бројн / Александра Бајазетов-Вучен	*Буриданов магарац*		2890
6	Кристоф Рансмајер / Златко Красни	*Последњи свет*		3107
7	Елена Феранте / Јелена Брборић	*Приче о онима који одлазе и онима који остају*	италијански	8316

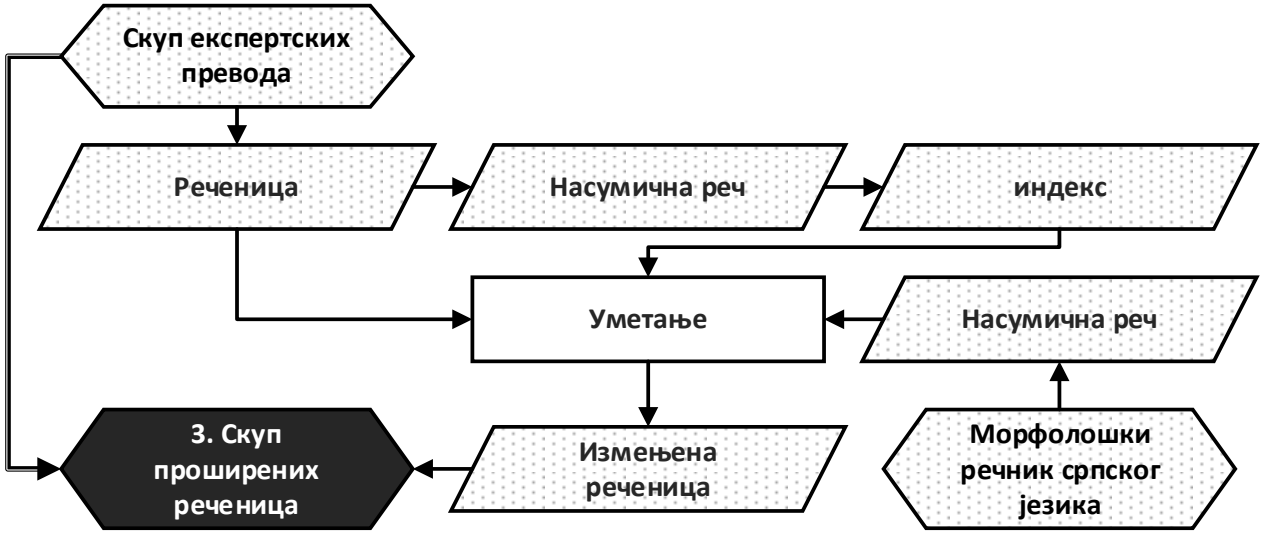
Први скуп података који ће се користити за евалуацију, јесте скуп реченица на српском језику из поменутих ресурса (скуп експертских превода), који ће се користити као репер при евалуацији. Ових реченица је издвојено укупно 26118.

Други, трећи и четврти скуп података добијени су применом једноставних алгоритама у комбинацији (за два од три) са морфолошким речником српског језика (Krstev, 2008; Stanković, et al., 2018). Као предуслов је, за сваку реченицу из првог скупа (скупа експертских превода), најпре насумично одређен неки индекс i, мањи од укупног броја речи у реченици која се посматра, и издвојена је реч са тим индексом, тј. реч на тој позицији у реченици. Даље процесирање се обавља у складу са том издвојеном речи и скупом реченица који желимо да добијемо.

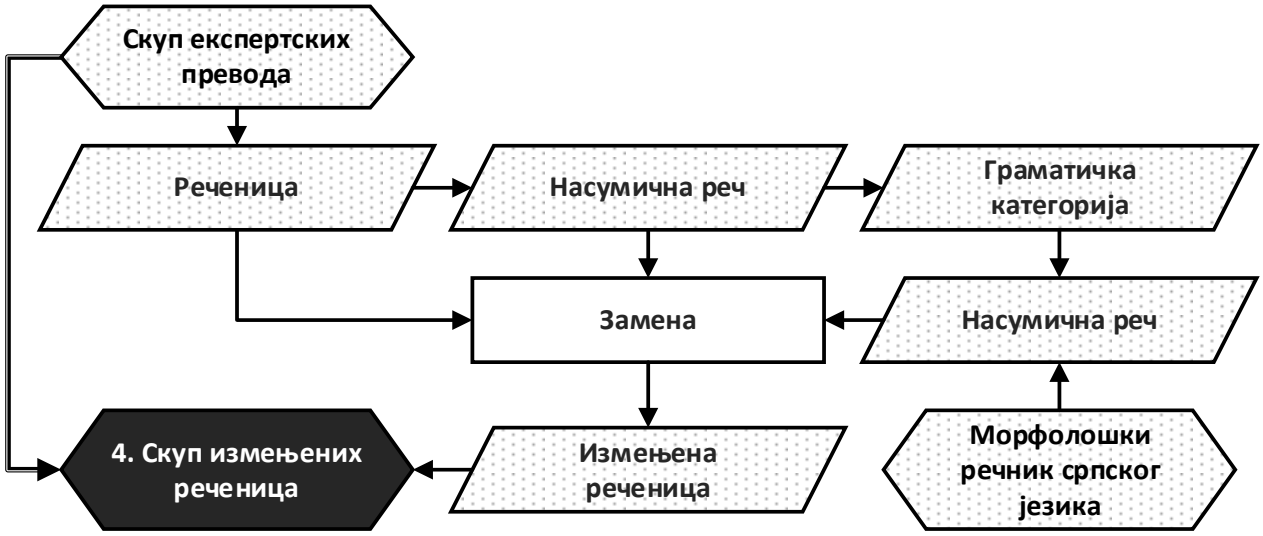
При креирању другог скупа (скупа окрњених реченица), издвојена реч је у свакој реченици једноставно уклоњена (Илустрација 40). У случају креирања трећег скупа (скупа проширених реченица) је пре издвојене речи (на њеном индексу) уметнута нова, насумична реч, односно флективни облик речи из морфолошког речника српског језика (Илустрација 41). За потребе креирања последњег скупа из ове групе, четвртог скупа (скупа измењених реченица), издвојена реч замењена је другом речи исте граматичке категорије из морфолошког речника: нпр. аниматна именица мушког рода у локативу једнине, замењује се другом речи истих граматичких својстава (Илустрација 42).



Илустрација 40: Креирање скупа окрњених реченица од скупа експертских превода.



Илустрација 41: Креирање скупа проширених реченица од скупа експертских превода уметањем насумичне речи односно флективног облика речи из морфолошког речника.



Илустрација 42: Креирање скупа измењених реченица од скупа експертских превода заменом одређени речи другом, али одговарајућом (према граматичкој категорији) из морфолошког речника.

Применом ових трансформација за, на пример, задати индекс 7 и реченицу „*Сећам се као да је било данас.*“, те узимањем насумичних речи из речника, добиле следеће три реченице:

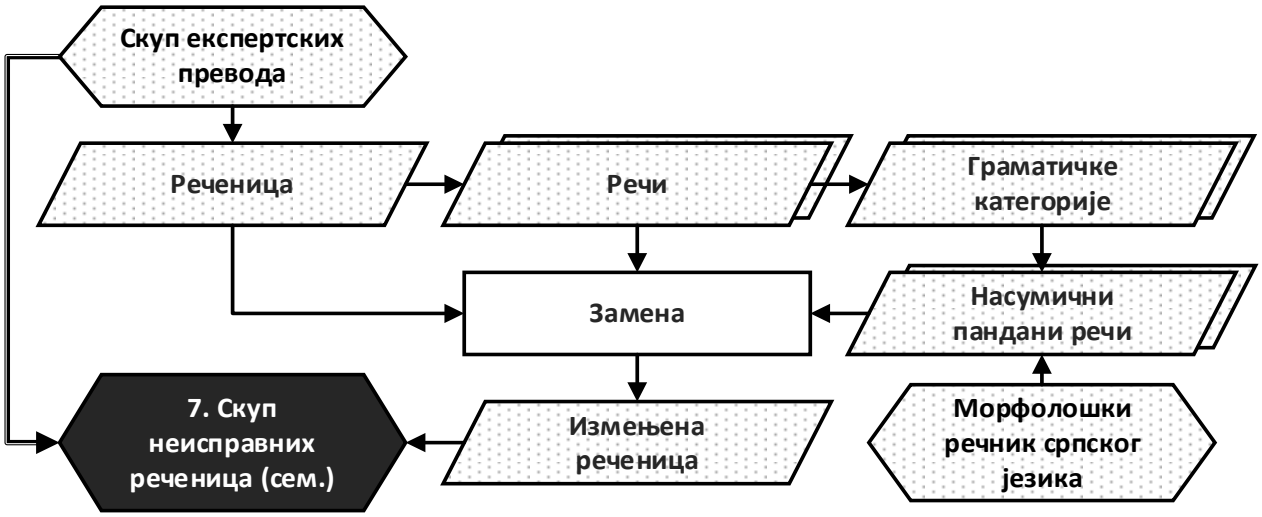
- *Сећам се као да је било.*
- *Сећам се као да је било* ***маса*** *данас.*
- *Сећам се као да је било* ***процветати****.*

Пети скуп података (први скуп синтаксички неисправних реченица) добијен је једноставном лематизацијом реченица из скупа експертских превода (Илустрација 43). Премда се може десити да је лематизована реченица једнака полазној (све речи у реченици су већ биле леме), једноставним поређењем једнакости између њих је израчунато да се то дешава у мање од 1% случајева.

Шести скуп података, који представља други скуп синтаксички неисправних реченица, добијен је изменом редоследа речи у реченицама из првог скупа (Илустрација 43). Као и у претходном случају, то не мора нужно значити да су реченице неисправне, али је ручном евалуацијом на скупу од 300 реченица установљено да се то, поново, дешава у мање од 1% случајева.

Илустрација 43: Креирање скупова синтаксички неисправних реченица.

Седми скуп података (скуп реченица исправних синтаксички, али не и семантички), добијен је заменом свих речи у реченици њиховим панданима (према граматичкој категорији) из морфолошког речника српског језика. Наиме, свака реч у реченици из скупа експертских превода се замењује другом, насумичном речи исте граматичке категорије по узору на алгоритам припремљен за креирање четвртог скупа, само што се овом приликом замењују све могуће речи, а не само по једна у свакој реченици (Илустрација 44).

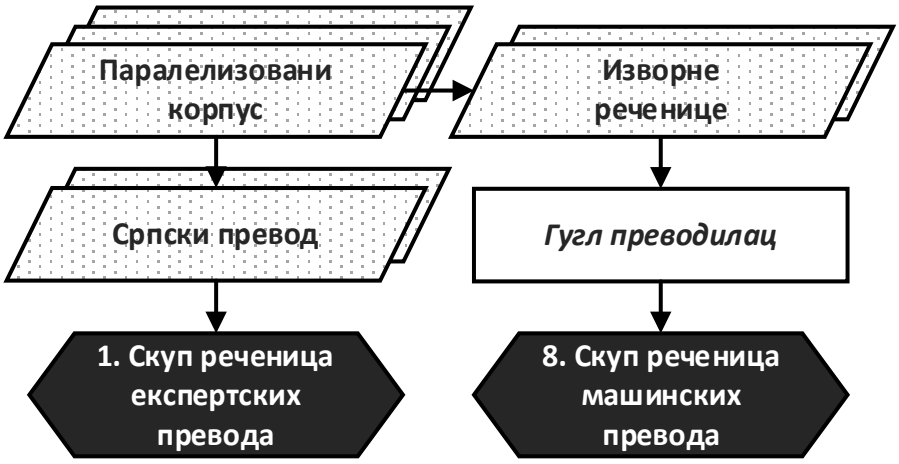


Илустрација 44: Креирање скупа семантички неисправних реченица од скупа експертских превода, заменом појединачних речи насумичним (исте граматичке категорије) из морфолошког речника.

Примери семантички неисправних реченица из скупова 5 (реченице неисправне према облику речи) и 6 (реченице неисправне према редоследу речи), као семантички неисправне реченице из скупа 7, генерисаних на основу претходно поменуте реченице били би:

- ***сећати*** *се као да је* ***бити*** *данас.*
- ***као данас било Сећам се да је.***
- ***Растрљавам*** *се као да је било данас.*

Последњи скуп за евалуацију добијен је преводом изворних реченица на немачком и италијанском језику из прикупљених ресурса на српски коришћењем сервиса *Гугл преводилац* (Илустрација 45).



Илустрација 45: Креирање скупа експертских и скупа машинских превода од паралелизованог корпуса уз помоћ сервиса Гугл преводилац.

Применом овог алгоритма добијен је паралелни корпус реченица које су преведене од стране експерта и реченица које су машински преведене[18], тј. први и осми скуп за евалуацију. Анализом њиховог садржаја утврђено је да се они разликују у око 97.77% случајева, а разлике су понекад и упечатљиве, на пример, исту реченицу коју је *Гугл* преводилац превео као:

Бесрамно је искористио неодлучност.

експертски преводилац је превео као:

Он је бесрамно искористио предност човека неспособног да донесе одлуку.

Задатак моделовања мини-језика и наведен уз овај скуп (Табела 23) се пре свега односи на могућност моделовања ових реченица, тј. класификације између ова два типа превода.

Сваки од осам креираних скупова је трансформисан у свој семантички и синтаксички пандан коришћењем трансформација описаних у поглављима 7.2.1 и 7.2.2, како би се директно могли евалуирати предобучени семантички и синтаксички модели, као и композитни модели који их користе.

9.2 Процес евалуације и резултати

За потребе евалуације модела креираних током овог истраживања најпре је извршена обрада свих скупова података (наведених у одељку 9.1) свим могућим самосталним и композитним моделима (описаним у одељцима 7.3, 8.2, 8.4, 8.5 и 8.6), тј. коришћењем сваког модела су израчунати скаларна перплексност са једне стране и вектор перплексности са друге стране, и то за сваку реченицу сваког скупа. Све перплексности су потом конвертоване у вероватноће, а добијене вредности су послужиле као основа евалуације модела на сва три предвиђена задатка, при чему је сваки задатак евалуиран на одређеним скуповима (Табела 23) и коришћењем одређених композиција, чија је припрема описана у одељку 8.

9.2.1 Детекција уклоњене, уметнуте и замењене речи

Приликом евалуације модела на задатку детекције уклоњених, уметнутих и замењених речи коришћени су вектори перплексности добијени кроз обраду реченица из другог, трећег и четвртог припремљеног скупа, дакле:

- Листе окрњених реченица, где је насумично уклоњена једна реч (на одређеном индексу);

[18] Управо из тог разлога су и коришћене реченице изворно писане на европским језицима јер би, у супротном, добили парове реченица од којих је једна писана изворно на српском језику а друга два пута преведена, што не би било погодно за поређење.

- Листе проширених реченица, где је (на истом индексу) насумично додата једна реч из речника;
- Листе реченица где је (на истом индексу) реч насумично замењена одговарајућом алтернативом из речника.

Свака реченица је обрађена коришћењем како самосталних (основни, семантички и синтаксички) модела (припремљених у одељку 7.3), тако и двадесет композитних модела описаних у одељку 8.4.1. Циљ модела је био да за сваку од реченица погоди спецификовани индекс (место где је реч уклоњена, уметнута или замењена).

За сваки модел и за сваку реченицу индекс је одабиран као индекс са најнижом мером вероватноће вектора перплексности реченице који је добијен помоћу модела у питању. Додатно је, као основица, послужио и метод насумичног одабира индекса за сваку реченицу. Овом приликом коришћене само реченице дуже од седам речи, што је дужина за два већа од задатог прозора коришћеног при креирању вектора, као што је појашњено у одељку 8.3. Реченица за тестирање је било укупно 8188.

Циљ евалуације био је да се издвоји модел или композиција који на овом задатку показују највећу стопу тачности (за сваки скуп појединачно). Мерена је тачност сваког модела при погађању индекса, при чему је сваки погодак утицао на повећање мере тачности која је израчуната као:

$$\text{тачност} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0, & a_i \neq b_i \\ 1, & a_i = b_i\end{cases}$$

где је *n* укупан број реченица дужих од седам речи, *n=8188*, *a* листа индекса највеће перплексности за вектор сваке од тих реченица и *b* листа индекса на којој је свака од тих реченица модификована.

Као алтернатива, из разлога што није једнако лако погодити индекс у реченицама различите дужине, рачуната је и мера тачности нормализована у односу на дужину, где се сваки погодак рачунао као разлика броја 2 и рециprочне дужине реченице (тако да би погодак на реченици дужине 1 вредео 1(само хипотетички, јер се користе само реченице дуже од седам речи), а погоци на дужим реченицама вредели више од тога):

$$\text{нормализована тачност} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0, & a_i \neq b_i \\ 2 - 1/l, & a_i = b_i\end{cases}$$

где је *n* укупан број реченица дужих од седам речи (8188), *a* листа индекса највеће перплексности за вектор сваке од тих реченица, *b* листа индекса на којој је свака од тих реченица модификована, а *l* дужина те реченице.

Резултати евалуације приказани су у наставку (Табела 25) и укључују и резултате које је постигао насумични одабир који је послужио као основица за упоређивање.

Табела 25: Резултати евалуације насумичног одабира, три самостална и двадесет композитних модела на задатку детекције места у реченици одакле је реч уклоњена, где је реч уметнута, или на коме је једна реч насумично замењена другом из речника (скупови 2, 3 и 4) Највећа тачност (међу самосталним и композитних моделима засебно) приказана је подебљано.

модел	скуп 2	скуп 3	скуп 4	скуп 2	скуп 3	скуп 4
	тачност			нормализована тачност		
насумично	0.0580	0.0312	0.0202	0.1114	0.0600	0.0387
основни	**0.1037**	**0.1726**	**0.1856**	**0.2000**	**0.3339**	**0.3593**
семантички	0.0526	0.0364	0.0303	0.1010	0.0698	0.0580
синтаксички	0.0760	0.0678	0.0602	0.1458	0.1299	0.1152
д-просек $_{0+1}$	0.1046	**0.1742**	0.1868	0.2017	**0.3369**	0.3616
д-производ $_{0+1}$	0.0367	0.0410	0.0410	0.0702	0.0787	0.0787
д-минимум $_{0+1}$	0.0555	0.0661	0.0679	0.1065	0.1269	0.1305
д-максимум $_{0+1}$	**0.1047**	0.1739	0.1868	**0.2019**	0.3365	0.3616
д-внорма $_{0+1}$	**0.1047**	**0.1742**	**0.1870**	**0.2019**	**0.3369**	**0.3619**
д-просек $_{0+2}$	0.0897	0.0819	0.0734	0.1725	0.1574	0.1409
д-производ $_{0+2}$	0.0807	0.1050	0.1070	0.1552	0.2021	0.2059
д-минимум $_{0+2}$	0.0936	0.1486	0.1580	0.1803	0.2871	0.3053
д-максимум $_{0+2}$	0.0843	0.0765	0.0681	0.1620	0.1469	0.1306
д-внорма $_{0+2}$	0.0863	0.0782	0.0697	0.1660	0.1502	0.1338
д-просек $_{1+2}$	0.0769	0.0693	0.0619	0.1475	0.1328	0.1185
д-производ $_{1+2}$	0.0463	0.0466	0.0455	0.0887	0.0891	0.0870
д-минимум $_{1+2}$	0.0555	0.0661	0.0679	0.1065	0.1269	0.1304
д-максимум $_{1+2}$	0.0762	0.0684	0.0609	0.1462	0.1312	0.1167
д-внорма $_{1+2}$	0.0762	0.0684	0.0609	0.1462	0.1312	0.1167
д-просек $_{0+1+2}$	0.0899	0.0821	0.0735	0.1728	0.1577	0.1411
д-производ $_{0+1+2}$	0.0361	0.0401	0.0401	0.0693	0.0769	0.0769
д-минимум $_{0+1+2}$	0.0555	0.0661	0.0679	0.1065	0.1269	0.1304
д-максимум $_{0+1+2}$	0.0843	0.0765	0.0681	0.1620	0.1469	0.1306
д-внорма $_{0+1+2}$	0.0863	0.0782	0.0697	0.1660	0.1502	0.1338

Најпре треба истаћи да резултати тачности и нормализоване тачности показују високу корелацију (преко 99%) у виду Пирсоновог коефицијента корелације:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

где је n величина узорка (низа), x и y популације вредности (тачности и нормализоване тачности), $\bar{x}$ и $\bar{y}$ аритметичке средине тих популација, а x_i и y_i елементи низа.

Даље, из приказаних резултата се види да од самосталних модела највећу тачност на овом задатку испољава основни модел, синтаксички модел има нешто слабију тачност, а семантички је најлошији. Основни модел увелико надмашује резултате насумичног одабира (од 78% до 827%), синтаксички модел такође (од 30% до 198%), док семантички модел не показује унапређења. Такође, види се да је најлакше детектовати замењену реч (18.56% тачности), па уметнуту (17.26% тачности), док је најтеже детектовати уклоњену реч (10.37% тачности).

Што се тиче композитних модела, само три (сва три заснована на комбинацији основног и семантичког модела) су успела да надмаше поменуте перформансе на сва три скупа: *д-просек* $_{0+1}$, *д-максимум* $_{0+1}$ и *д-внорма* $_{0+1}$. Од њих, највећу тачност на сва три скупа постигао је композитни модел *д-внорма* $_{0+1}$ за који су измерене тачности од 10.47%, 17.42% и 18.70%, што одговара побољшањима од око 1% понаособ у односу на основни модел.

9.2.2 Детекција семантичких и синтаксичких неправилности

Приликом евалуације модела на задатку детекције семантичких и синтаксичких неправилности коришћени су резултати скаларне перплексности добијене при евалуацији друга три припремљена скупа података:

- Листа реченица које нису синтаксички исправне, добијених лематизацијом свих речи;
- Листа реченица које нису синтаксички исправне, добијених насумичним мешањем речи;
- Листа реченица које су исправне синтаксички, али не и семантички, добијена насумичном заменом речи.

Циљ је био установити које су разлике у вероватноћама када се посматрају реченице које су узете за златни стандард (скуп 1) у односу на вероватноће добијене обрадом синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7), а тестирано је три обучена самостална језичка модела (припремљена у одељку 7.3) и 120 композитних псеудограматика (из одељака 8.2 и 8.4.2).

На пример, добар модел за откривање семантички неисправних реченица би за сваки пар реченица (од којих је једна експертски превод а друга њен пандан из скупа 7), за

прву увек давао уверљиво вишу вероватноћу, а код идеалног модела се распони вредности вероватноћа реченица из првог и седмог скупа не би уопште преклапали.

Дакле, детекција неисправности се огледала у поређењу израчунате вероватноће реченице из првог скупа и њеног парњака из петог, шестог или седмог, што је представљало три одвојена задатка. За сваки од три задатка и за сваки модел које се евалуира израчунате су две мере: просечна удаљеност вредности вероватноћа и максимална тачност.

Мера просечне удаљености вредности (у даљем тексту удаљеност) је израчуната на основу процентуалне површине пресека хистограма вероватноћа за реченице из првог (*a*) и другог скупа (*b*). Хистограми су дефинисани на основу 500 одељака вероватноће (*h*). Дакле, све вредности од 0 до 1 распоређене су у групе са маргином прецизности мањом од 0.002. Површина пресека хистограма израчуната је као сума разлика веће и мање популације за сваки одељак хистограма, подељена са двоструким укупним бројем реченица (*n*), а удаљеност се рачуна као разлика броја један и добијене површине:

$$\text{удаљеност} = 1 - \frac{1}{2n}\sum_{i=1}^{n}\left||a_i - b_i|\right|$$

Максимална тачност је израчуната као разлика броја један и количника минималне грешке при класификацији (*False positive* + *False Negative, FP+FN*) и двоструког укупног броја реченица:

$$\text{тачност} = 1 - \frac{1}{2n}\sum_{i=1}^{n}\min(FP_i + FN_i)$$

Насумичним одабиром, а с обзиром да је ово двокласна класификација, добила би се тачност од 0.5, тј. 50%, а сваки резултат већи од тога је позитиван. Ипак, идеалан модел би имао што већу израчунату вредност удаљености и што већу максималну тачност. Резултати евалуације за три самостална модела приказани су испод (Табела 26), а резултати евалуације за двадесет основних композитних модела (описаних у одељку Пример 8.2) приказани су у наставку текста (Табела 27).

Табела 26: Резултати евалуације три самостална језичка модела на задатку детекције синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7).

модел	скуп 5		скуп 6		скуп 7	
	удаљ.	тачност	удаљ.	тачност	удаљ.	тачност
основни	**0.6557**	**0.8286**	**0.7703**	**0.8870**	**0.6045**	**0.8025**
семантички	0.0344	0.5003	0.0859	0.5428	0.2924	0.6477
синтаксички	0.3050	0.6519	0.3688	0.6843	0.0395	0.5010

Табела 27: Резултати евалуације двадесет основних композитних модела на задатку детекције синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7).

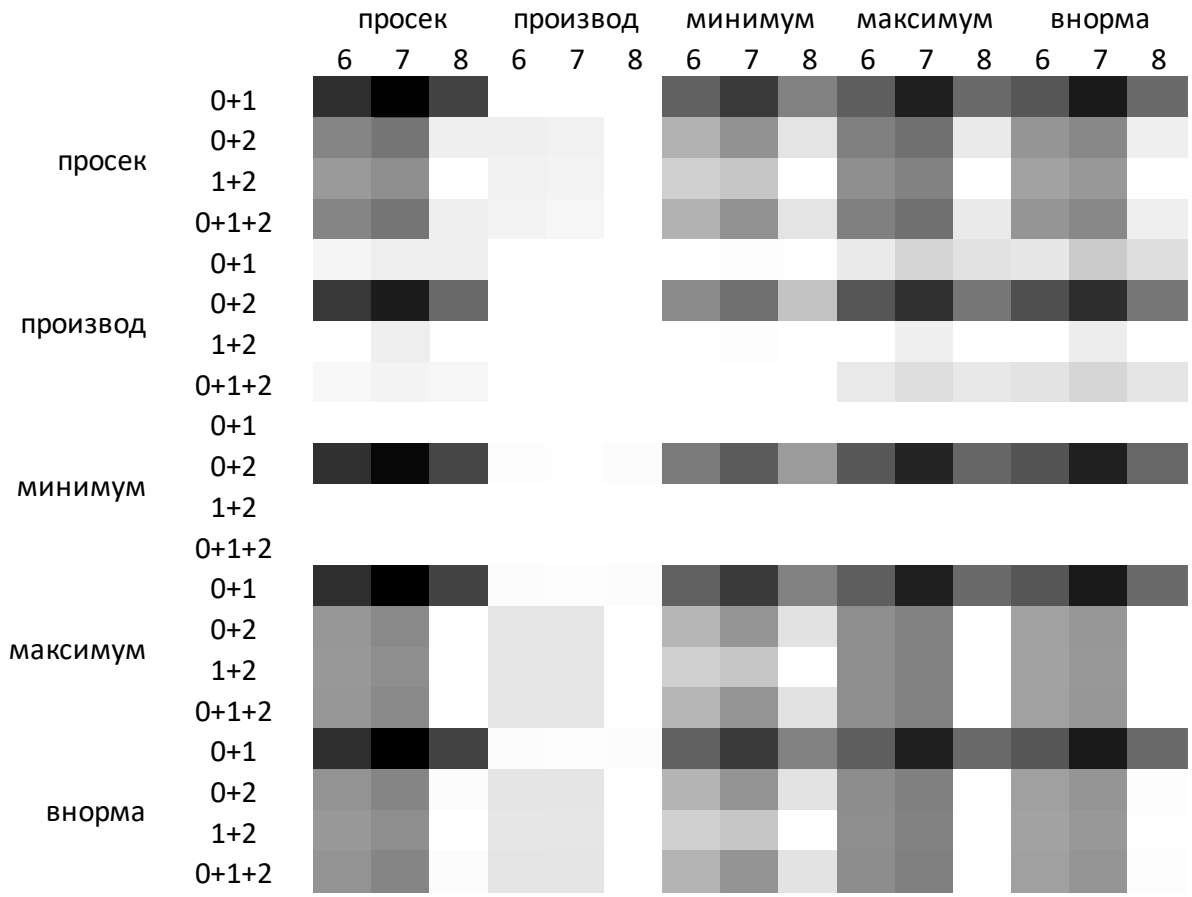
модел	скуп 5		скуп 6		скуп 7	
	удаљ.	тачност	удаљ.	тачност	удаљ.	тачност
к-просек $_{0+1}$	0.3206	0.6612	0.4642	0.7329	0.5362	0.7688
к-производ $_{0+1}$	0.1339	0.6603	0.1452	0.7372	0.1444	0.7254
к-минимум $_{0+1}$	0.3454	0.6736	0.4930	0.7467	0.3488	0.6758
к-максимум $_{0+1}$	0.3236	0.6614	0.4427	0.7218	0.5889	0.7952
к-внорма $_{0+1}$	0.3290	0.6641	0.4613	0.7305	0.5700	0.7856
к-просек $_{0+2}$	0.3927	0.6966	0.4669	0.7337	0.0869	0.5389
к-производ $_{0+2}$	0.6374	0.8242	0.7292	0.8677	0.5011	0.7515
к-минимум $_{0+2}$	**0.6523**	**0.8266**	**0.7644**	**0.8839**	**0.5986**	**0.7995**
к-максимум $_{0+2}$	0.3053	0.6521	0.3695	0.6845	0.0452	0.5041
к-внорма $_{0+2}$	0.3219	0.6597	0.3888	0.6942	0.0555	0.5084
к-просек $_{1+2}$	0.2914	0.6459	0.3759	0.6881	0.0537	0.5193
к-производ $_{1+2}$	0.1268	0.5639	0.2371	0.6233	0.2396	0.6244
к-минимум $_{1+2}$	0.0333	0.5060	0.1018	0.5511	0.2897	0.6464
к-максимум $_{1+2}$	0.3040	0.6513	0.3679	0.6838	0.0440	0.5025
к-внорма $_{1+2}$	0.3028	0.6510	0.3710	0.6853	0.0515	0.5046
к-просек $_{0+1+2}$	0.3805	0.6905	0.4698	0.7350	0.1267	0.5596
к-производ $_{0+1+2}$	0.0141	0.6977	0.0136	0.7675	0.0137	0.7142
к-минимум $_{0+1+2}$	0.3459	0.6740	0.4922	0.7463	0.3466	0.6745
к-максимум $_{0+1+2}$	0.3043	0.6514	0.3685	0.6840	0.0485	0.5066
к-внорма $_{0+1+2}$	0.3180	0.6588	0.3921	0.6963	0.0637	0.5158

Из приказаних резултата се види да семантички модел показује веома лоше резултате при детекцији синтаксичких грешака, са просечном удаљеношћу од 3.44% и 8.59%, док је тачност тек нешто виша од насумичне (50.03% и 54.28%). Насупрот томе, он остварује знатно вишу удаљеност, као и тачност од 64.77% при детекцији семантички неисправних реченица. Исто тако, синтаксички модел остварује добре резултате при детекцији синтаксичких грешака (тачност од 65.19% и 68.43% и удаљеност од 30.5% и 36.88%) док при детекцији семантичких грешака има веома ниску тачност—50.1% и удаљеност мању од 4%. Ипак, према свим критеријума и за све задатке, најбоље

резултате остварује поново основни модел, са удаљеностима од 65.57%, 77.03% и 60.45%, као и тачностима у висини од 82.86%, 88.7% и 80.25%.

У складу са тиме, ни резултати добијени евалуацијом основних композитних модела не показују побољшање. Најбољи модел на свим задацима је комбинација основног и синтаксичког модела, *к-минимум* $_{0+2}$. Како синтаксички модел уобичајено даје највишу перплексност, метод композиције коришћењем минимизације своди ову композицију мање-више на емулацију основног модела и производи резултате у складу са тиме: удаљености од 65.23%, 76.44% и 59.86%, те тачности од 82.66%, 88.39% и 79.95%.

Потрага се наставља анализом композиција заснованих на сажетим векторима вероватноћа (описаних у одељку 8.4.2). Како тих модела има укупно сто и евалуирају се помоћу две мере и на три различита задатка, резултати су, зарад прегледности, приказани у виду топлотне мапе. Мапа је сложена тако да наглашава најподобније комбинације начина комбиновања (просек, производ, минимум, максимум и внорма) по редовима и сажимања (истих пет) по колонама, те унутрашњу поделу по коришћеним моделима (0, 1 и 2) и скуповима за евалуацију (6, 7 и 8), при чему се посматра метрика тачности (Илустрација 46).



Илустрација 46: Топлотна мапа постигнутих максималних тачности композитних модела заснованих на векторима вероватноћа, при чему су модели груписани према начину комбиновања вектора (по редовима) и сажимања добијеног вектора у скалар (по колонама). Тамнија боја означава већи проценат тачности.

Из илустрације се наслућује да метод свођења није толико битан колико је битно који се модели користе за компоновање, али и очигледно је да је коришћење производа при свођењу лоша идеја, што је и интуитивно, јер у том случају вероватноћа готово у потпуности зависи од дужине реченице (експоненцијално се смањује за сваки нови токен).

Укупно најбоља композиција од свих из ове групације је *к-максимум-просек* $_{0+1}$ која остварује удаљености од 59.6%, 72.86% и 53,61% и тачности од 79.85%, 86.4% и 76.9%, што су резултати лошији и од оних које су постигли самостални модели (Табела 26), као и од оних које су постигле основне композиције (Табела 27).

Други сет тестова који је спроведен у оквиру овог испитивања имао је за циљ да евалуира могућност модела да праве разлику између реченица које су неисправне синтаксички и реченица које су неисправне семантички. Поново су рачунате мере удаљености и максималне тачности, али овога пута су посматрани парови вероватноћа реченица у скуповима 5 и 7 и скуповима 6 и 7 (у оба теста је један скуп реченица синтаксички, а други семантички неисправан). Резултати самосталних модела на ова теста приказани су испод (Табела 28).

Табела 28: Резултати евалуације три самостална језичка модела на задатку разликовања синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7). Тест 1 приказује разлику између семантички неисправних реченица и оних синтаксички неисправних из скупа 5, док тест 2 приказује разлику између семантички неисправних реченица и синтаксички неисправних реченица из скупа 6.

модел	тест 1		тест 2	
	удаљеност	тачност	удаљеност	тачност
основни	0.1117	0.5559	**0.4386**	**0.7192**
семантички	**0.3174**	0.5103	0.2261	0.5206
синтаксички	0.3123	**0.656**	0.3787	0.6889

Из приказаних резултата се види да, за разлику од оних где су тражене разлике између исправних и неисправних реченица, на овим тестовима резултати нису толико једнострани у корист основног модела. Премда семантички модел и даље показује максималну тачност тек нешто вишу од насумичне, он овога пута показује највишу меру удаљености на првом од два теста. Синтаксички модел овог пута показује боље резултате од основног модела по свим параметрима на првом тесту (побољшање тачности од преко 17% у односу на основни модел) док на другом тесту показује добре, али нешто лошије резултате у односу на њега (4% лошија максимална тачност).

Исти тестови спроведени су и на композитним моделима (двадесет основних композиција и сто заснованих на сведеним векторима вероватноћа). Резултати тих тестова приказани су у наставку текста (Табела 29).

Табела 29: Резултати евалуације двадесет основних композитних модела на задатку разликовања синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7) Тест 1 приказује разлику између семантички неисправних реченица и оних синтаксички неисправних из скупа 5, док тест 2 приказује разлику између семантички неисправних реченица и синтаксички неисправних реченица из скупа 6.

модел	тест 1		тест 2	
	удаљеност	тачност	удаљеност	тачност
к-просек $_{0+1}$	0.2279	0.5037	0.1737	0.5495
к-производ $_{0+1}$	0.0114	0.5079	0.0020	0.5539
к-минимум $_{0+1}$	0.0377	0.5119	0.2472	0.6270
к-максимум $_{0+1}$	0.2722	0.5028	0.2399	0.5499
к-внорма $_{0+1}$	0.2530	0.5032	0.2148	0.5504
к-просек $_{0+2}$	**0.3183**	**0.6596**	0.3954	0.6977
к-производ $_{0+2}$	0.2387	0.6414	0.3305	**0.7565**
к-минимум $_{0+2}$	0.1172	0.5587	**0.4396**	0.7196
к-максимум $_{0+2}$	0.3123	0.6560	0.3787	0.6889
к-внорма $_{0+2}$	0.3143	0.6563	0.3803	0.6897
к-просек $_{1+2}$	0.2593	0.6296	0.3449	0.6721
к-производ $_{1+2}$	0.1133	0.5154	0.0362	0.5432
к-минимум $_{1+2}$	0.3005	0.5104	0.2137	0.5209
к-максимум $_{1+2}$	0.3081	0.6538	0.3739	0.6865
к-внорма $_{1+2}$	0.2992	0.6494	0.3686	0.6840
к-просек $_{0+1+2}$	0.2633	0.6317	0.3630	0.6815
к-производ $_{0+1+2}$	0.0005	0.5161	0.0003	0.6155
к-минимум $_{0+1+2}$	0.0369	0.5124	0.2482	0.6276
к-максимум $_{0+1+2}$	0.3080	0.6538	0.3739	0.6865
к-внорма $_{0+1+2}$	0.2999	0.6495	0.3710	0.6847

И овога пута најбоље резултате показују комбинације основног и синтаксичког модела, али сада они надмашују резултате основног модела (и семантичког тамо где је он најбољи). Композиција *к-просек* $_{0+2}$ на првом тесту надмашује тачност основног модела за 18.7% (и синтаксичког за мање од 1%), док композиција *к-производ* $_{0+2}$ надмашује основни модел за 15.3% на првом тесту и 5.1% на другом тесту, а лошији је

од синтаксичког модела на првом тесту за мање од 1%, а надмашује га на другом тесту за 9.8%.

Што се тиче резултата композиција заснованих на сведеним векторима они поново, у великој већини показују лошије резултате, са изузетком композиције *к-максимум-максимум* $_{0+1+2}$ која показује свеукупно највишу тачност на првом тесту — 66.75%, што је побољшање од 1.7% у односу на најбољи основни композитни модел, али на другом тесту показује лошије резултате (тачност од 68.56%).

Коначно, како би се тестирала уопштена моћ детекције неисправних реченица, направљен је додатни скуп података, који се састоји од трећине реченица из скупова 5, 6 и 7, дакле, мешовити скуп неисправних реченица различитог порекла. Тестирање је поновљено коришћењем основних и композитних модела, а резултати су приказани у табели испод (Табела 30).

Табела 30: Резултати евалуације три самостална и двадесет основних композитних модела на задатку детекције неисправних реченица мешовитог порекла.

модел	удаљеност	тачност
основни	**0.6646**	**0.8324**
семантички	0.1182	0.5592
синтаксички	0.2171	0.6082
к-просек $_{0+1}$	0.4353	0.7186
к-производ $_{0+1}$	0.1407	0.7046
к-минимум $_{0+1}$	0.3804	0.6909
к-максимум $_{0+1}$	0.4521	0.7264
к-внорма $_{0+1}$	0.4542	0.7270
к-просек $_{0+2}$	0.3104	0.6553
к-производ $_{0+2}$	0.5849	0.8008
к-минимум $_{0+2}$	0.6554	0.8284
к-максимум $_{0+2}$	0.2206	0.6095
к-внорма $_{0+2}$	0.2441	0.6217
к-просек $_{1+2}$	0.2341	0.6171
к-производ $_{1+2}$	0.2012	0.6050
к-минимум $_{1+2}$	0.1264	0.5635
к-максимум $_{1+2}$	0.2244	0.6111

к-внорма $_{1+2}$	0.2228	0.6109
к-просек $_{0+1+2}$	0.3212	0.6609
к-производ $_{0+1+2}$	0.0141	0.7232
к-минимум $_{0+1+2}$	0.3812	0.6911
к-максимум $_{0+1+2}$	0.2197	0.6093
к-внорма $_{0+1+2}$	0.2452	0.6216

Из резултата се види да најбољу могућност генерализације поново има основни модел, што је највероватније последица лоших резултата које постижу семантички и синтаксички модел, па самим тим и већина композиција које их укључују.

9.2.3 Моделовање мини-језика

Тестови тачности и удаљености су поновљени и за задатак моделовања мини-језика, тј. за разлучивање експертских од машинских превода. У овом случају, тражени су модели који би просечно веће вероватноће додељивали експертским преводима (скуп 1) у односу на машинске преводе (скуп 8), међутим ниједан од модела није томе тежио. Резултати евалуације за самосталне језичке моделе, основне композиције и пробране композиције (шест најбољих) засноване на сажетим векторима приказани су испод (Табела 31).

Табела 31: Резултати евалуације три самостална и двадесет основних композитних модела на задатку детекције машинских превода (скуп 8).

модел	удаљеност	тачност
основни	0.0393	0.5023
семантички	0.0412	0.4999
синтаксички	**0.0854**	**0.5024**
к-просек $_{0+1}$	0.0441	0.5004
к-производ $_{0+1}$	0.0227	0.5000
к-минимум $_{0+1}$	0.0281	0.4999
к-максимум $_{0+1}$	0.0435	0.5005
к-внорма $_{0+1}$	0.0426	0.5005
к-просек $_{0+2}$	0.0820	0.5030
к-производ $_{0+2}$	0.0613	**0.5026**
к-минимум $_{0+2}$	0.0378	0.5023

к-максимум $_{0+2}$	**0.0866**	**0.5026**
к-внорма $_{0+2}$	0.0854	**0.5026**
к-просек $_{1+2}$	0.0759	0.5017
к-производ $_{1+2}$	0.0400	0.4999
к-минимум $_{1+2}$	0.0369	0.4999
к-максимум $_{1+2}$	0.0872	0.5018
к-внорма $_{1+2}$	0.0898	0.5018
к-просек $_{0+1+2}$	0.0803	0.5020
к-производ $_{0+1+2}$	0.0084	0.5000
к-минимум $_{0+1+2}$	0.0274	0.4999
к-максимум $_{0+1+2}$	0.0871	0.5020
к-внорма $_{0+1+2}$	0.0879	0.5020
к-минимум-максимум $_{0+1}$	0.0026	**0.5134**
к-минимум-внорма $_{0+1}$	0.0038	0.5131
к-минимум-максимум $_{1+2}$	0.0030	**0.5134**
к-минимум-внорма $_{1+2}$	**0.0043**	0.5131
к-минимум-максимум $_{0+1+2}$	0.0026	**0.5134**
к-минимум-внорма $_{0+1+2}$	0.0038	0.5131

Из резултата се може видети да тачност за све моделе и композиције ординира око 50% (±1.34%), што указује на то да ниједан не решава овај задатак.

9.2.4 Евалуација композиција заснованих на вештачким неуронским мрежама

Задатак наслаганих перцептрона је био да потврде успешност композитних модела која је претходно истражена (одељци 5 и 6), али на задацима детекције семантички и синтаксички неисправних реченица, разлика између њих као и моделирања мини-језика и детекције неисправних реченица уопште. За задатке из прве групације (два детекције синтаксички и један детекције семантички неисправних реченица) је било потребно да са статистичком значајношћу надмаше основни модел, док је за задатке друге групације (разликовање синтаксички и семантички неисправних реченица, као и детекција неисправних реченица мешовитог порекла) био је неопходно надмашити најбољи композитни модел. За моделирање мини-језика било потребно само да са статистичком значајношћу реше проблем (статистички значајан резултат са тачношћу од преко 50%).

Као што је већ поменуто у одељку 8.6, обучено је укупно пет перцептрона (на одељцима од по 80% реченица из одговарајућих скупова), који су потом на преосталих 20% евалуирани на поменута четири задатка, а резултати су упоређени са онима које је постигао основни језички модел.

Резултати свих обучених перцептрона (у виду мере тачности) приказани су заједно са онима које је на истом исечку остварио основни језички модел на четири дефинисана задатка у наставку текста (Табела 32).

Табела 32: Резултати које су остварile композиције засноване на наслаганом перцептрону, упоређене са резултатима које је остварио основни модел на задацима детекције синтаксички неисправних реченица (скупови 5 и 6), семантички неисправних реченица (скуп 7) и машинских превода (скуп 8).

модел	скуп 5	скуп 6	скуп 7	скуп 8
основни	0.8286	0.8870	0.8025	0.5023
к-перцептрон-1	0.8519	0.8921	0.7970	0.5319
к-перцептрон-2	0.8479	0.8886	0.7986	0.5328
к-перцептрон-3	0.8653	0.8954	0.8132	0.5329
к-перцептрон-4	0.8627	0.8919	0.8137	0.5306
к-перцептрон-5	0.8456	0.8874	0.7988	0.5337
к-перцептрон просек	**0.8547**	**0.8911**	**0.8042**	**0.5324**

Резултати наслаганих перцептрона на задатку разлучивања синтаксички и семантички неисправних реченица, упоређени су са најбољим пређашњим резултатима у табели испод (Табела 33).

Табела 33: Резултати које су остварile композиције засноване на наслаганом перцептрону, упоређене са резултатима које је остварио најбољи пређашњи модел на задацима разликовања синтаксички (скупови 5 и 6) и семантички неисправних реченица (скуп 7), као и неисправних реченица мешовитог порекла.

модел	скуп 5 и скуп 7	скуп 6 и скуп 7	мешовити скуп
најбољи модел до сада	0.6596	0.7565	**0.8324**
к-перцептрон-1	0.7257	0.7760	0.8328
к-перцептрон-2	0.7165	0.7870	0.8155
к-перцептрон-3	0.7140	0.8040	0.8061
к-перцептрон-4	0.7266	0.7950	0.8148
к-перцептрон-5	0.7255	0.7932	0.7533
к-перцептрон просек	**0.7217**	**0.7910**	0.8045

Из приказаних резултата се најпре види композиције засноване на наслаганим перцептронима надмашују резултате које је остварио основни модел на прва четири задатка (Табела 33), при чему је на задатку детекције синтаксички неисправних реченица (према облицима речи) остварено највеће побољшање (3.15%), где је измерен степен смањења грешке од 18%, док су на задатку детекције синтаксички неисправних реченица (према редоследу речи у реченици) остварена побољшања од мање од 1% тј. смањења грешке од 3.67%. На задатку детекције семантички неисправних реченица су пронађена побољшања која нису статистички значајна, док је на задатку моделовања мини-језика тј. детекције машинских превода остварена просечна тачност од 53.24%, најбољи резултат до тада.

Са друге стране, на задацима разликовања синтаксички и семантички неисправних реченица (Табела 33), ова композиција је показала додатна, значајна, побољшања: на првом задатку остварено је побољшање од 9.4% у односу на композицију *к-просек* $_{0+2}$ и побољшање од 10.1% у односу на најбољи самостални модел, док је на другом задатку остварено побољшање тачности од 4.56% у односу на композицију *к-производ* $_{0+2}$, односно 9.98% у односу на најбољи самостални модел.

Што се тиче резултата остварених на задатку детекције неисправних реченица из мешовитог скупа ниједна композиција није забележила побољшање у односу на основни језички модел. Ту, као и на задатку моделовања мини-језика примећено је недовољно уклапање (*underfitting*) тј. примећено је да модел није могао бити обучен до потребне мере, стога су на овим задацима даље тестиране композиције засноване на конволуционим неуронским мрежама (описане у одељку 8.6), које имају већу моћ моделовања[19]. Резултати ових испитивања приказани су у наставку текста (Табела 34).

Табела 34: Резултати које су остварile композиције засноване на наслаганој конволуционој неуронској мрежи, упоређене са резултатима које је остварио најбољи пређашњи модел на задацима детекције машинских превода тј. моделовања мини језика (скуп 8) и неисправних реченица мешовитог порекла.

модел	скуп 8	мешовити скуп
најбољи модел до сада	0.5324	0.8324
к-конволуција-1	0.5525	0.9234
к-конволуција-2	0.5473	0.9167
к-конволуција-3	0.5500	0.9195
к-конволуција-4	0.5443	0.9265
к-конволуција-5	0.5373	0.8875
к- конволуција просек	**0.5462**	**0.9147**

[19] Упркос томе што нису остварена изузетна побољшања на прва три теста, параметри обучавања указивали су на то да је скуп за обучавање адекватно генерализован и постигнута је минимална ентропија, што би значило да употреба дубоког учења у овом случају не би донела побољшање.

Из приложених резултата је очигледно да је употреба оваквог класификатора на овим задацима била оправдана, а пре свега на задатку детекције неисправних реченица из мешовитог скупа где је тачност побољшана у просеку за чак 9.89% у односу на основни језички модел (13.7% у односу на наслагани перцептрон) уз смањење грешке од скоро 50%. Што се тиче задатка моделовања мини-језика и ту је у свакој итерацији остварено побољшање тачности, са просечним побољшањем од 2.6% у односу на наслагани перцептрон и 8.72% у односу на најбољи самостални језички модел.

9.3 Евалуација генеративних граматика

Евалуација генеративних граматика осмишљених за потребе овог истраживања заснивала се на евалуацији реченица које оне генеришу (пример: Табела 22). Коришћењем основног језичког модела (генеративни предобучени трансформер) обученог за потребе овог истраживања (одељак 7.3), као и седам додатних композитних метода генерисања (одељак 8.5, Табела 21), припремљено је, за ту потребу, укупно осам корпуса од по хиљаду машински генерисаних реченица. Сви модели су добили исти заметак тј. исту почетну реченицу — *Сећам се као да је било данас.* Она је послужила као контекст за генерисање следеће, прве машински генерисане реченице, а потом је свака следећа реченица генерисана коришћењем њене претходнице као контекста.

Све реченице сваког корпуса евалуиране су коришћењем најбољих произведених метода евалуације. За потребе овог тестирања обучена су четири нова композитна модела (при чему је архитектура бирана на основу претходно приказаних резултата), од којих је сваки скројен тако да испитује одређено својство генерисаног текста. Задаци на којима су реченице тестиране, тј. својства која су испитивана, модели помоћу којих су својства евалуирана и скупови на којима су ти модели обучени приказани су у наставку текста (Табела 35).

Табела 35: Својства генерисаних реченица која се испитују, модели који се користе за њихову евалуацију и скупови на којима су ти модели обучени.

својство које се испитује	модел за евалуацију	скупови на коме је модел обучен (прва класа, друга класа)
синтаксичка исправност (према облицима речи)	е-перцептрон-синт1	скуп 1 (100%), скуп 5 (100%)
синтаксичка исправност (према редоследу речи)	е-перцептрон-синт2	скуп 1 (100%), скуп 6 (100%)
семантичка исправност	е-перцептрон-сем	скуп 1 (100%), скуп 7 (100%)
општа исправност	е-конволуција	скуп 1 (100%), скупови 5, 6 и 7 (по ~33.33%)

За сваки креирани корпус, коришћењем четири поменута модела за евалуацију, израчуната је просечна вероватноћа реченица:

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

где је n укупан број реченица тј. n=1000 и P_i вероватноћа неке специфичне реченице, Након тога израчуната је и стандардна девијација вероватноће на скупу вероватноћа реченица, и то помоћу претходно добијене просечне вероватноће:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(P_i - \bar{P})^2}{n}}$$

Циљ је био да се пронађе композитни модел који производи реченице које су највероватније, док је стандардна девијација уведена као помоћна мера која пружа увид у његову стабилност. Резултати просечне вероватноће за свих осам корпуса, и сва четири модела приказани су у табели испод (Табела 36), док су резултати стандардне девијације приказани у наставку текста (Табела 37).

Табела 36: Резултати исправности реченица машински генерисаних помоћу различитих метода (редови), изражене у виду просечних вероватноћа израчунатих према различитим критеријумима (колоне), тј. коришћењем различитих модела за евалуацију, при чему последња колона одражава просечну вероватноћу према осталим критеријумима.

модел	синт. исправност (облици)	синт. исправност (редослед)	сем. исправност	општа исправност	просек
основни	0.9510	0.9666	0.9469	0.7302	0.8987
г 0+1	**0.9812**	0.9867	**0.9791**	**0.7311**	**0.9195**
г 0+2	0.9772	0.9874	0.9752	0.7310	0.9177
г-просек 0+1+2	0.9749	0.9843	0.9733	0.7310	0.9159
г-производ 0+1+2	0.9811	0.9870	0.9785	**0.7311**	0.9194
г-минимум 0+1+2	0.9763	0.9860	0.9754	0.7284	0.9166
г-максимум 0+1+2	0.9799	**0.9885**	0.9783	0.5965	0.8858
г-внорма 0+1+2	0.9777	0.9872	0.9767	**0.7311**	0.9181

Табела 37: Стандардне девијације вероватноћа реченица машински генерисаних помоћу различитих метода (редови), израчунатих према различитим критеријумима (колоне), при чему последња колона одражава просечну девијацију тј. просечну вредност осталих колона.

модел	синт. исправност (облици)	синт. исправност (редослед)	сем. исправност	општа исправност	просек
основни	0.1091	0.0907	0.1128	0.0120	0.0812
г $_{0+1}$	0.0707	0.0624	0.0760	**0.0000**	0.0523
г $_{0+2}$	0.0716	0.0520	0.0721	0.0004	**0.0490**
г-просек $_{0+1+2}$	0.0809	0.0636	0.0809	0.0002	0.0564
г-производ $_{0+1+2}$	**0.0679**	0.0600	0.0738	**0.0000**	0.0504
г-минимум $_{0+1+2}$	0.0798	0.0588	0.0781	0.0173	0.0585
г-максимум $_{0+1+2}$	0.0671	**0.0518**	**0.0690**	0.0590	0.0617
г-внорма $_{0+1+2}$	0.0755	0.0565	0.0742	**0.0000**	0.0515

Из добијених резултата се види да већина композитних модела за генерисање надмашује самостални модел, штавише он на готово свим задацима показује најлошије резултате и при мерењу просечне вероватноће и при мерењу стандардне девијације (изузетак је само тест опште исправности где је лошији резултат на оба мерења показао композитни модел *г-максимум $_{0+1+2}$*).

Ако пак посматрамо моделе који су се најбоље показали понајвише се истиче композитни модел *г $_{0+1}$* (реченице се генеришу помоћу основног модела, а најбољег кандидата бира семантички модел), који је остварио највишу просечну вероватноћу на три од четири задатка, укључујући онај можда најбитнији, тестирање опште исправности. На том задатку је остварио и стандардну девијацију нижу од 0.00005, што говори и о његовој стабилности. Укупно просечно побољшање остварено у односу на основни модел је 2.31%.

Најбоље и најлошије оцењене (према просечној вероватноћи) реченице у сваком од скупова приказане су у наставку текста (Табела 38)

Табела 38: Просечно најбоље и најлошије оцењена реченица у сваком од осам генерисаних скупова.

модел	Најбоље оцењена реченица	Најлошије оцењена реченица
основни	*Изложбу је отворио Милорад Екмечић, градоначелник Новог Сада.*	*У браку са Стравинским, познао је Годуајев изузетно интелигентни и склони љубав према музици.</ср.<тв.<С.М. Магазин/Ма*

г 0+1	*Он је указао да се Србија противи да се изручење Васиљковића оцени као кршење међународних права и да је протекло више од годину дана од доношења пресуде у том суду.*	*дивизијске словенске дивизе прожети армијом капитала.*
г 0+2	*У сукобима је повређено 12 полицајаца, а рањени су полицајци, ватрогасци, припадници војске, жандармерије и жандармерије.*	*Хердерова оставка, која је добила већину резултата из првог дела, била је потпуна.<сола је замењена претњом шефом државе.<садашњи министар правде и пред*
г-просек 0+1+2	*У ствари сам био професор историје уметности на Архитектонском факултету у Београду.*	*226.<сијецам Иназада, Абидада Иназда је преузео власт над Абадом и прикључио се Ахабади, који је касније постао претендент на престо.<т*
г-производ 0+1+2	*Иначе, у влади Србије су се у последњих неколико дана појавиле различите примедбе на начин на који ће бити формулисани предлози закона о министарствима.*	*Он се приближава, али не и старим схватањем и наглашава своје скромно осећање задовољства које није ствар времена већ осећања среће.<тине љубави.<<схва*
г-минимум 0+1+2	*Судије су оцениле да је нелогично да тужиоци завршe са испитивањем, и затражиле покретање прекршајног поступка.*	*и једини елемент географије, универзални и једини спољашњи облик.*
г-максимум 0+1+2	*Ипак, ова мисија је почела да ради у децембру прошле године.*	*атистике-слободног кретања производних податаке.*
г-внорма 0+1+2	*Прецизније речено, интересовање ђака за упис прве генерације оснивача и средњошколаца за основно образовање на Косову и Метохији.*	*Због тога што су замрзнуте грађанске ратове и укидање робног промета, стари кући су постали део државе.</сног града.</л<среу.<т излажу на истој*

10
Дискусија

Основна претпоставка на којој се овај рад заснива је да се на основу пробабилистичких излаза више различитих језичких модела, било заснованих на различитим технологијама или на различитим репрезентацијама текста (на пример, текст у коме су речи сведене на лему или граматичку категорију) примењених над анотираним текстом, може направити композитни, интелигентни систем чије ће перформансе надмашивати перформансе тих модела употребљених појединачно. Та претпоставка успешно потврђена на иницијалним експериментима у обради српског језика (Stanković, et al., 2022; Škorić, et al., 2022), чиме је потврђен и потврдан одговор на прва три истраживачка питања којима се овај рад бави (одељци 5 и 6):

ИП1 Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и њиховим излазним вероватноћама адекватна метода у области обраде природних језика?

ИП2 Да ли се композитни интелигентни системи засновани на паралелним моделима могу надограђивати – и да ли већи број активних паралелних модела побољшава квалитет целокупног система?

ИП3 Да ли паралелни језички модели боље моделирају језик од појединачних?

Након тога, приступљено је, најпре, раду на утврђивању корпуса који ће се користити за коначни експеримент. Финални корпус за ово истраживање је тако скројен од постојећег корпуса квалитетног српског језика, *СрпКор2013* (који је обезбедило *Друштво за језичке ресурсе и технологије* и који чини отприлике 50% финалног корпуса), потом дела његове најновије допуне, корпуса *СрпКор2021,* који је припремљен, између осталог, за потребе овог рада (и чини око 20% финалног корпуса) и, коначно, јавних корпуса српског језика, *ВикиКорпус* и *СрпЕЛТеК*, начињених од текстова са Википедије на српском језику и старих српских романа (који чине отприлике 30% финалног корпуса).

Једном припремљен и уједначен, корпус је употребљен за обучавање савремених језичких модела, генеративних предобучених трансформера друге генерације (*GPT-2*). У трајању од две седмице, обучена су три модела, од којих један контролни (основни) и два експериментална (семантички и синтаксички) која су обучавана на специјално припремљеним репрезентацијама корпуса, што је описано у одељку 7. Први је обучен на корпусу обрађеном методама латентне семантичке анализе, при чему су уклоњене све стоп речи (речи које нису именице, глаголи, придеви, прилози или бројеви) док је остатак лематизован, у циљу да добијени текст испољава искључиво семантички аспект текста. Други модел је обучаван на корпусу који је обрађен коришћењем морфолошких речника српског језика и тагера који обележава прецизне граматичке категорије (на пример: именица мушког рода у номинативу множине, аниматна). Где год је то било могуће, речи су замењене граматичким категоријама, а циљ је био да се добијени текст лиши значења и да представља синтаксно могуће реченице српског језика.

Коначно, како би се одговорило на преостала три истраживачка питања

ИП4 Да ли композитне псеудограматике боље моделирају језик од појединачних?
ИП5 Који су оптимални методи комбинације излаза појединачних језичких модела?
ИП6 Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и излазним вероватноћама подобна метода у области интелигентних система уопште?

припремљени језички модели комбиновани су на основу својих пробабилистичких излаза, те коришћењем различитих архитектура, у композитне псеудограматике које су потом тестиране на скупу реченица претходно некоришћених за обучавање. Скројено је укупно 149 различитих композиција од којих се 122 користе за евалуацију квалитета, 20 за проналажење грешака у тексту и 7 за генерисање нових реченица (одељак 8). Перформансе ових композиција евалуиране су, дакле, на посебно припремљеним корпусима на укупно десет задатака, у циљу међусобног поређења. Три задатка се односе на детекцију грешака у тексту

- детекција избрисане речи у тексту;
- детекција уметнуте речи у тексту;
- детекција замењене речи у тексту;

док се преосталих седам односе на евалуацију њиховог квалитета (било у служби евалуације или генерисања):

- детекција семантички неисправних реченица;
- детекција синтаксички неисправних реченица (два задатка);
- разлучивање семантички и синтаксички неисправних реченица (два задатка);
- детекција неисправних реченица уопштено;
- разлучивање између експертских и машинских превода.

10.1 Постигнути резултати

Евалуација добијених система на неколико припремљених задатака описана је у одељку 9, а коришћене су претежно методе аутоматске евалуације засноване на интринсичној мери перплексности. Већ резултати постигнути на задацима који се тичу детекције грешака у тексту тј. детекције уметнуте, уклоњене и замењене речи (Табела 25) указивали су на супериорност контролног језичког модела, обученог на основном, неизмењеном корпусу квалитетног текста. Он је том приликом показао знатно боље перформансе од синтаксичког модела, док је семантички модел остварио резултате тек нешто боље од насумичних. Ипак, треба напоменути да је он (семантички модел) због природе предобраде коју користи (уклањање стоп речи) био у приличном хендикепу на овим задацима, јер је насумични индекс који треба погодити могао указивати на реч којој он није ни имао приступ. Упркос томе, композитни модел заснован на комбинацији основног и семантичког модела остварио је најбоље резултате према свих шест посматраних критеријума (тачност и нормализована тачност на задацима погађања уметнуте, уклоњене и замењене речи), мада, за малу маргину од око 1%.

Када су у питању задаци који се тичу евалуације реченица, постигнути резултати су били слични, где је у највећем броју случајева најбољи модел (од самосталних) био управо основни (Табела 26, Табела 28, Табела 30), а композиције засноване на правилима (укључујући и оне засноване на векторима перплексности) су давале мала унапређења (Табела 29). Ови иницијални резултати указали су на важност обучавања, пре свега, додатног синтаксичког модела (који је показао добре резултате на задатку разазнавања између семантички и синтаксички неисправних реченица, и то поготово у комбинацији са основним моделом), као и корисност композиција заснованих на паралелним језичким моделима уопште.

Највеће унапређење је пак (као и у иницијалним експериментима описаним у одељцима 5 и 6) дошло са коришћењем наслаганих класификатора заснованих на хеуристикама. Композитни модели засновани на наслаганом перцептрону (Илустрација 37) су показали следеће перформансе, приказане у виду постотка увећања тачности и смањења грешке при класификацији (Табела 39):

Табела 39: Процентуална побољшања у виду повећања тачности и смањења грешке које су остварили композиције засноване на наслаганом перцептрону у односу на најбољи самостални модел на седам задатака.

задатак	проценат повећања тачности	проценат смањења грешке
детекција синтаксички неисправних реченица (према облицима речи)	3.15	18.00
детекција синтаксички неисправних реченица (према редоследу речи)	0.46	3.67

детекција семантички неисправних реченица	0.21	0.86
разликовање синтаксички и семантички неисправних реченица (први тест)	10.10	23.61
разликовање синтаксички и семантички неисправних реченица (други тест)	9.98	28.22
детекција неисправних реченица уопштено	/	/
разликовање експертских и машинских превода	6.00	6.44

Осим што за задатак детекције уопштено неисправних реченица нису остварена побољшања, приликом тестирања установљено је да се за последња два задатка модели не обучавају до довољне мере (*underfitting*), па су стога за њих специјално обучени нешто комплекснији наслагани класификатори у виду конволуционих неуронских мрежа (Илустрација 39), које као улазне податке користе векторе перплексности, тј. векторе вероватноће (описане у одељку 8.3). Резултати које су остварили ови модели приказани су испод (Табела 40).

Табела 40: Процентуална побољшања у виду повећања тачности и смањења грешке које су остариле композиције засноване на наслаганој конволуционој неуронској мрежи у односу на најбољи самостални модел на два најтежа задатака.

задатак	проценат повећања тачности	проценат смањења грешке
детекција неисправних реченица уопштено	9.89	48.24
разликовање експертских и машинских превода	8.74	9.67

Даље, при евалуацији модела у служби генерисања нађено је да свих седам предложених композиција надмашују резултате (премда су измерене разлике мале) које остварује основни језички модел (Табела 36), при чему свеукупно најбоље резултате остварује комбинација основног и синтаксичког језичког модела илити проверавање синтаксичке исправности генерисаних кандидата.

Ови резултати сами по себи дају (а поготово када се сагледају остварења на последња два задатка) несумњив позитиван одговор на ИП4:

Да ли композитне псеудограматике боље моделирају језик од појединачних?

Поред тога, говоре нам и то да су наслагани класификатори засновани на хеуристикама (а поготово они засновани на векторима перплексности и конволуционим неуронским мрежама) оптималан метод композиције, као одговор на ИП5:

Који су оптимални методи комбинације излаза појединачних језичких модела?

10.2 Примена и значај

Као што је већ установљено, у раду се разматра проблем моделирања језика, где се посебна пажња посвећује коришћењу већег броја расположивих ресурса, при чему се *ресурс* не односи само на корпус текстова, већ и друге, претходно припремљене језичке ресурсе (попут морфолошких речника) или различите технологије моделирања. Имајући у виду њихов растући број, композитни модел је изразита прилика за њихово обједињавање и заједничку употребу.

Током истраживања остварени су очекивани научни доприноси у виду:

- **Проширења корпуса савременог српског језика**. Проширење се огледа у групном раду на објављивању најновије допуне корпуса савременог српског језика, *СрпКор2021* (описано у одељку 7.1.2), као и корпуса старих српских романа, *СрпЕЛТеК* (описаног у одељцима 6.4.1 и 7.1.3).
- **Развијања новог софтвера за аутоматску анотацију корпуса, на основу прикупљених предобележених примера**. Овај софтвер (описан у одељку 5), заснован је на паралелној употреби предобучених модела за анотацију и приликом евалуације је остварио најбоље резултате међу свим тестираним системима за српски језик. Софтвер је написан у програмском језику Пајтон, а код је у слободном приступу доступан на платформи ГитХаб (*GitHub*)[20].
- **Развијања савремених језичких модела српског језика на основу различитих репрезентација корпуса савременог српског језика**. За потребе истраживања обучено је три савремена језичка модела српског језика у виду предобучених генеративних трансформера друге генерације тј. ГПТ-2 (описано у одељку 7.3), а обучавања су вршена у складу са трансформацијама текста (описано у одељку 7.2) заснованим на верзији корпуса проширеном (помоћу претходно поменутог софтвера за анотацију) врстама речи, ширим граматичким категоријама и лемама. Развијени модели су отворени за јавност и доступни на сајту заједнице *huggingface*[21], као и путем Пајтон пакета *transformers* под истим именима.
- **Развијања детаљног модела композитног система за паралелно обједињавање креираних модела (укључујући и будуће моделе) и креирање псеудограматика српског језика које ће имати примену у задацима обраде природног језика, укључујући класификацију и евалуацију докумената, као и генерисање текста**. Током истраживања је скројено 149 различитих композитних система од којих се 122 користе за евалуацију квалитета (120 заснованих на правилима и описаних у одељцима 8.2 и 8.4.2, и два заснована на хеуристикама, описана у одељку 8.6), 20 се користе за проналажење грешака у тексту (описано у одељку 8.4.1) и 7 за генерисање нових реченица (описано у одељку 8.5). Софтвер који демонстрира генерисање текста на основу креираних модела, те евалуацију унетог текста коришћењем припремљених граматика,

[20] https://github.com/procesaur/BEaSTagger, приступљено 20. јануара 2023.

[21] https://huggingface.co/procesaur/**gpt2-srlat**, https://huggingface.co/procesaur/**gpt2-srlat-sem**, https://huggingface.co/procesaur/**gpt2-srlat-synt**, приступљено 20. јануара 2023.

као и његову трансформацију у векторе перплексности, припремљен је у виду веб апликације писане у програмском језику Пајтон, чији је код такође доступан у слободном приступу на платформи ГитХаб (*GitHub*)[22], а поменута веб апликација је доступна и на вебу[23].

- **Тестирања и валидације на постојећим, претходно истраженим проблемима**. Све креиране композиције тестиране су на познатим проблемима попут детекције неисправних реченица и њиховој класификацији.

Развијене псеудограматике ће наћи, дакле, примену у решавању различитих проблема у обради природног језика, јер се, попут формалних граматика, и оне могу користити у морфосинтаксној анализи текста, за генерисање новог текста, као и за прецизније израчунавање сличности између текстова или евалуацију њиховог квалитета.

Једном усавршена псеудограматика ће, дакле, поред генеративног апарата, представљати и систем за евалуацију који са статистичком значајношћу и на континуираној скали одређује квалитет текста који му је дат на улазу. Такав систем би сам по себи имао неколико примена, као што је већ наведено у секцији 3.2. Такође, с обзиром на то да су машинско превођење, аутоматско генерисање текста и одговарање на упите већ умногоме аутоматизовани, одговарајућа аутоматска евалуација би додатно побољшала, а уз адекватну примену, и убрзала и олакшала те задатке. Осим тога, аутоматско проналажење (и потенцијално исправка) неквалитетног текста била би још једна могућа примена оваквог система, под условом да се он додатно усаврши.

Највећи допринос у смислу новитета огледа се у обучавању синтаксичких и семантичких генеративних језичких модела, као и у употреби мере перплексности при њиховој комбинацији. Синтаксички језички модел је, пре свега, показао велику дискриминациону вредност на задатку разликовања синтаксички и семантички неисправних реченица, али и играо улогу у успеху композитних модела, док је семантичком моделу на изглед потребна дорада. Њиховим унапређењем путем додатних фиих подешавања процедура за предобраду би се резултати могли додатно унапредити.

Што се тиче комбиновања на нивоу одаслате перплексности, она омогућава композицију не само модела исте бранше, већ било којих који производе перплексност или вероватноће, независно од типа и димензија, па би, на пример, омогућила комбинације *GPT* и *BERT* модела, или чак комбинације језичких модела са онима који нису језички.

Још један новитет који се уводи, а који би такође могао бити широко примењив јесу вектори перплексности, који се такође, путем процедура приказаних у овом раду,

[22] https://github.com/procesaur/Parallel-language-models, приступљено 20. јануара 2023.

[23] https://plma.jerteh.rs, приступљено 20. јануара 2023.

могу генерисати коришћењем било којег модела који испољава перплексност или вероватноћу за неки унос. Њихова прва (и показано оправдана употреба) лежи у детекцији грешака у тексту. Премда могућности нису потпуно истражене током приказаних експеримената, њихова комбинација са конволуционим неуронским мрежама показала се као пун погодак, при чему су створени неки од најбољих композитних модела, а који би се могли на лак начин додатно проширити са додатним *BERT* моделом, који би додао нову димензионалност карактеристикама које неуронска мрежа користи.

Наравно, сваки додатни модел који се примењује над текстом повећава време процесирања, у овом случају линеарно, када је придодата евалуација коришћењем семантичког (чија је примена незнатно бржа) и синтаксичког модела (чија је примена незнатно спорија у односу на основни језички модел). Даље, израчунавање вектора перплексности је спорије од израчунавања скаларне перплексности, а интензитет успорења зависи од дужине уноса и величине прозора: просечно време израчунавања скаларне перплексности у овом експерименту је око 0.14 секунди, док је израчунавање вектора перплексности негде око једне секунде, дакле око седам пута спорије. Када је у питању компоновање ових резултата оно траје незнатно у односу на време евалуације, чак и када је у питању обрада коришћењем конволуционе неуронске мреже, која је, ипак, неколико редова величине мања од најмањег ГПТ-2 модела. Додатно успорење дешава се и при генерисању текста где се време извршења линеарно повећава сразмерно броју кандидата који се генеришу од стране основног модела.

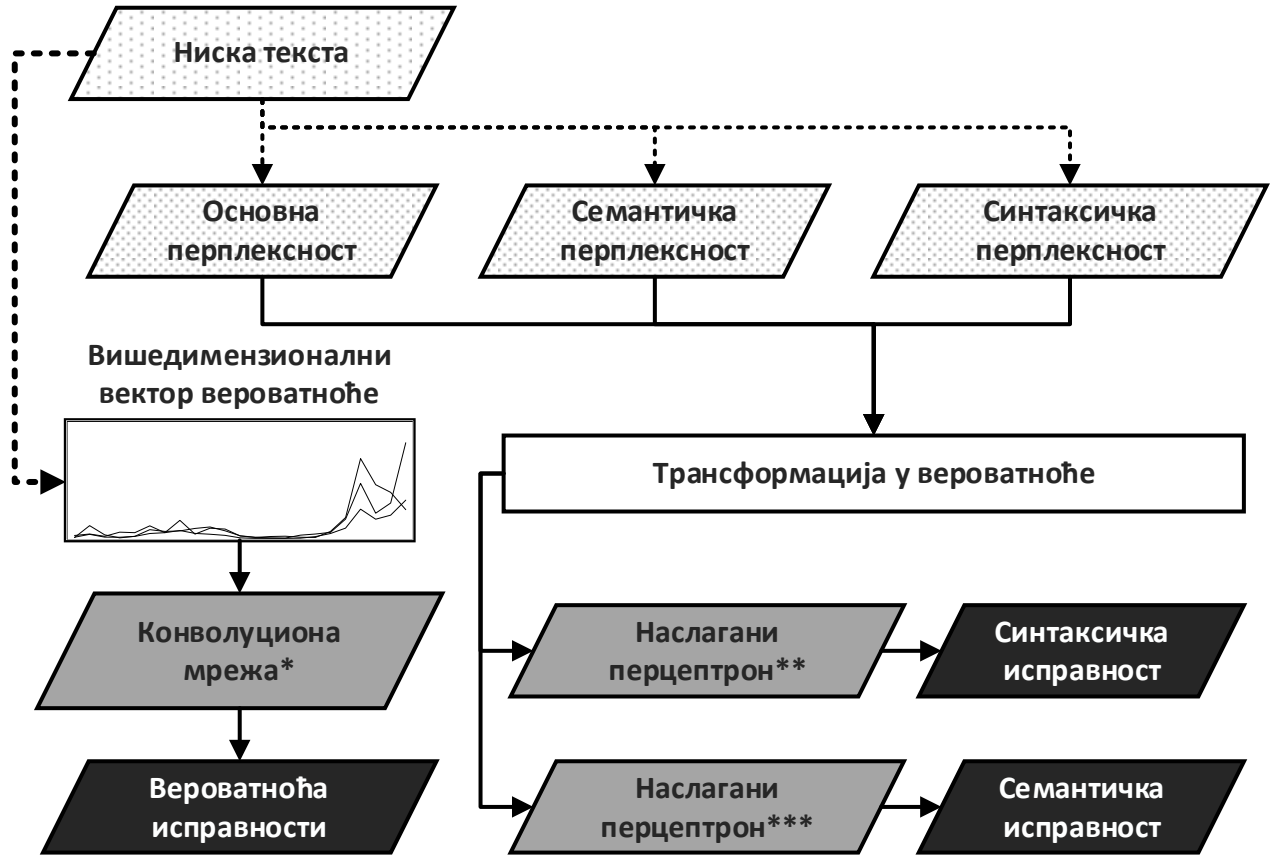
Оно што је такође значајно је што овај рад предлаже систем довољно флексибилан да се може проширивати и прилагођавати новим, нетестираним, употребама. Због тога ће знање прикупљено током истраживања, осмишљене архитектуре за повезивање, као и резултати евалуације тих архитектура омогућити развој нових или обједињавање постојећих интелигентних система, који не морају имати везе са обрадом природног језика. Успешност ове архитектуре (а поготово успешност њених различитих варијација) остварена на приказаним задацима индукује позитиван одговор и на ИП6:

Да ли је коришћење композитних интелигентних система заснованих на паралелним моделима и излазним вероватноћама је подобна метода у области интелигентних система уопште?

У употребљеним методама, као и у примени, огледа се мултидисциплинарност овог истраживања. Коришћене методе припадају различитим областима науке, а пре свега области обраде природног језика, која је и сама мултидисциплинарна, и где ће развијене методе и развијени ресурси наћи и највећу примену, поготово уз пригодну експлоатацију.

10.3 Закључак

На основу приказаних резултата закључујемо да је употреба композитних система заснованих на паралелним (језичким) моделима и њиховим пробабилистичким производима адекватна метода у области обраде природних језика, а поготово српског језика, и то за задатак апроксимације његове формалне граматике. Из приложеног се види да композитне псеудограматике у великој већини случајева надмашују резултате које остварују генеративни језички модели, како на задатку евалуације и класификације (са смањењем степена грешке и до 48%), тако и на задатку генерисања текста. Даље, закључујемо да је најподобнији метод комбиновања излаза језичких метода наслагани класификатор заснован на хеуристикама и за евалуацију текста предлажемо следећу псеудограматику, засновану на израчунатим вредностима перплексности, векторима перплексности, те наслаганим перцептронима и конволуционим неуронским мрежама (Илустрација 47).



Илустрација 47: Предлог структуре псеудограматика у служби евалуације текста која би на основу унетог текста, коришћењем претходно описаних метода обраде и компоновања испољавала вероватноћу исправности уноса, као и степен његове синтаксичке и семантичке исправности. **е-конволуција **е-перцептрон-синт2 ***е-перцептрон-сем*

У случају генерисања текста, препоручујемо структуру псеудограматике засновану на генерисању кандидата помоћу основног језичког модела и потом одабира најбољег коришћењем синтаксичког језичког модела (Илустрација 48). Овај приступ се намеће

као оптимално решење, поготову у случајевима где је квалитет генерисаног текста пресуднији од времена потребног за извршавање.

Илустрација 48: Предлог структуре псеудограматике која би генерисала квалитетне реченице, засновано на *n* кандидата које предлаже основни језички модел и одабира најбољег кандидата коришћењем синтаксичког језичког модела.

Даљи рад на овом проблему ће се свакако састојати од додатог проширења корпуса квалитетног текста и обучавања бољих (већих) и квалитетнијих језичких модела (на пример обучавање синтаксичког модела над корпусом анотираним пуном синтаксном анотацијом у виду функције речи у реченици). Даље, радиће се на усавршавању процеса аутоматског обележавања са једне стране, и репрезентација корпуса заснованих на новообележеном корпусу са друге стране. Детаљније експериментисање са детекцијом уклоњених, уметнутих и замењених речи, при чему би се обратила посебна пажња на њихову врсту речи и урадила компаративна анализа, би могла да употпуни тај аспекат истраживања.

Такође, даље истраживање биће усмерено на побољшање композиција тј. на побољшање наслаганих класификатора (коришћењем алтернативних структура вештачких неуронских мрежа), као и на осмишљање нових комбинација које би могле да убрзају рад уз очување постигнутог учинка. Коначно, адаптацијом развијеног софтверског решења у веб сервисе омогућило би се да он постане компонента других система и самим тим би се омогућила и шира експлоатација развијених модела.

11
Библиографија

Abney, S., 1997. Part-of-speech tagging and partial parsing. У: *Corpus-based methods in language and speech processing.* s.l.:Спрингер, pp. 118--136.

Agarwal, A. & Lavie, A., 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. У: *Proceedings of the Third Workshop on Statistical Machine Translation.* s.l.:s.n., pp. 115--118.

Agichtein, E. и други, 2008. Finding high-quality content in social media. *08 Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 183-194.

Akimushkin, C., Diego R, A. & Osvaldo N, O. J., 2018. n the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications,* Том 495, pp. 49--58.

Aliwy, A. H., 2015. Combining POS taggers in master-slaves technique for highly inflected languages as Arabic. У: *2015 International Conference on Cognitive Computing and Information Processing (CCIP).* s.l.:s.n., pp. 1--5.

Andonovski, J., Šandrih, B. & Kitanović, O., 2019. Bilingual lexical extraction based on word alignment for improving corpus search. *The Electronic Library,* 37(4), pp. 722-739.

Antiqueira, L., Nunes, M. d. G. V., Oliveira, O. N. & Costa, L. d. F., 2007. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications,* 373(1), pp. 811-820.

Antiqueira, L., Nunes, M. d. G. V., Oliveira, O. N. & Costa, L. d. F., 2009. A complex network approach to text summarization. *Information Sciences,* 179(5), pp. 584-599.

Arefyev, N., Sheludko, B. & Aleksashina, T., 2019. Combining neural language models for word sense induction. У: *International Conference on Analysis of Images, Social Networks and Texts.* s.l.:s.n., pp. 105--121.

Banerjee, S. & Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. У: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* s.l.:s.n., pp. 65--72.

Belz, A. & Reiter, E., 2006. Comparing automatic and human evaluation of NLG systems. *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 313-320.

Bird, S., Klein, E. & Loper, E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* s.l.:O'Reilly Media, Inc..

Bisk, Y., Zellers, R., Gao, J. & Choi, Y., 2020. Piqa: Reasoning about physical commonsense in natural language. У: *Proceedings of the AAAI conference on artificial intelligence.* s.l.:s.n., pp. 7432--7439.

Broman, S. & Kurimo, M., 2005. *Methods for Combining Language Models in Speech Recognition.* Lisbon, Interspeech, pp. 1317-1320.

Brown, P. F. и други, 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics,* 18(1), pp. 31--40.

Brown, P. F. и други, 1992. Class-based n-gram models of natural language. *Computational linguistics,* 18(4), pp. 467--480.

Brown, T. и други, 2020. Language models are few-shot learners. *Advances in neural information processing systems,* Том 33, pp. 1877--1901.

Burrows, J., 2002. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing,* 17(3), pp. 267--287.

Carroll, J. & Long, D., 1989. *Theory of finite automata with an introduction to formal languages.* s.l.:s.n.

Chen, Z., Lee, K.-F. & Li, M.-j., 2000. Discriminative training on language model. У: *Sixth International Conference on Spoken Language Processing.* s.l.:s.n.

Chomsky, N., 1956. Three models for the description of language. *IRE Transactions on information theory,* 2(3), pp. 113-124.

Chomsky, N., 2013. Problems of projection. *Lingua,* Том 130, pp. 33-49.

Chomsky, N. & Schützenberger, M. P., 1959. The algebraic theory of context-free languages. *Studies in Logic and the Foundations of Mathematics,* Том 26, pp. 118--161.

Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L., 1989. Finite State Automata and Simple Recurrent. *Neural Computation,* Том 1, pp. 372-381.

Cohen, V. & Gokaslan, A., 2020. Opengpt-2: open language models and implications of generated text. *XRDS: Crossroads, The ACM Magazine for Students,* pp. 26--30.

Conneau, A. & Lample, G., 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems,* Том 32.

Csuhaj-Varjú, E., 1994. Grammar Systems: a Multi-Agent Framework for Natural Language Generation. У: *Mathematical aspects of natural and formal languages.* s.l.:s.n., pp. 63-78.

De Beaugrande, R., 1987. Special purpose language and linguistic theory. *Unesco Alsed-LSP Newsletter,* 10(2), pp. 1977-2000.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding.* s.l.:arXiv preprint arXiv:1810.04805.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. У: *Proceedings of the second international conference on Human Language Technology Research.* s.l.:s.n., pp. 138--145.

Dong, L. и други, 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems,* Том 32.

Đorđević, B., 2017. *Izrada osnova formalne gramatike srpskog jezika upotrebom metagramatike.* s.l.:Универзитет у Београду, Филолошки факултет.

Eder, M. & Górski, R., 2022. Stylistic Fingerprints, POS-tags and Inflected Languages: A Case Study in Polish. *arXiv preprint arXiv:2206.02208.*

Eder, M., Rybicki, J. & Kestemont, M., 2016. Stylometry with R: a package for computational text analysis. *The R Journal,* 8(1).

El Manar El Bouanan, S. & Kassou, I., 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications,* 86(12).

Elman, J. L., 1988. *CRL Tech. Rep. 9901,* San Diego, CA: Center for Research in Language, University of California.

Elman, J. L., 1990. Finding Structure in Time. *Cognitive science,* 14(2), pp. 179-211.

Evangelopoulos, N., 2013. Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science,* 4(6), pp. 683--692.

Evert, S. и други, 2015. Towards a better understanding of Burrows's Delta in literary authorship attribution. У: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature.* s.l.:s.n., pp. 79--88.

Floyd, R., 1962. Algorithm 97: shortest path. *Communications of the ACM,* 5(6), p. 345.

Gao, L. и други, 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Gao, Leo; Biderman, Stella; Black, Sid; Golding, Laurence; Hoppe, Travis; Foster, Charles; Phang, Jason; He, Horace; Thite, Anish; Nabeshima, Noa; others.*

Geman, S. & Johnson, M., 2002. "Probabilistic grammars and their applications. У: *International Encyclopedia of the Social & Behavioral Sciences.* s.l.:s.n., pp. 12075-12082.

Giles, C. L. и други, 1992. Learning and Extracting Finite State Automata with Second-Order Recurrent Neural Networks. *Neural Computation,* Том 4, pp. 393-405.

Hacioglu, K. & Ward, W., 2001. *On Combining Language Models: Oracle Approach.* San Diego, Association for Computational Linguistics, pp. 1-4.

Hardcastle, D. & Scott, D., 2008. Can we evaluate the quality of generated text?. *LREC*, May.

Harrison, M. A., 1978. *Introduction to formal language theory.* s.l.:Addison-Wesley Pub. Co.

Hauser, M. D., Chomsky, N. & Fitch, W. T., 2002. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?. *Science,* 298(5598), pp. 1569-1579.

Henrich, V., Reuter, T. & Loftsson, H., 2009. CombiTagger: A System for Developing Combined Taggers. У: *FLAIRS Conference.* s.l.:s.n.

Hochreiter, S., 1991. *Untersuchungen zu dynamischen neuronalen Netzen,* München: Technische Universität München.

Hochreiter, S. & Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation,* 9(8), pp. 1735-1780.

Iyer, A. & Vosoughi, S., 2020. Style Change Detection Using BERT. У: *CLEF (Working Notes).* s.l.:s.n.

Jager, G. & Rogers, J., 2012. Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 367(1598), pp. 1956--1970.

Joshi, M. и други, 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics,* Том 8, pp. 64--77.

Kalchbrenner, N., Grefenstette, E. & Blunsom, P., 2014. *A Convolutional Neural Network for Modelling Sentences,* s.l.: arXiv preprint arXiv:1404.2188.

Karttunen, L., Chanod, J.-P., Grefenstette, G. & Schille, A., 1996. Regular expressions for language engineering. *Natural Language Engineering,* 2(4), pp. 305--328.

Kilgarriff, A. и други, 2014. The Sketch Engine: ten years on. *Lexicography,* 1(1), pp. 7--36.

Kingma, D. P. & Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Klakow, D. & Peters, J., 2002. Testing the correlation of word error rate and perplexity. *Speech Communication,* 38(1-2), pp. 19--28.

Kleene, S., 1956. Representation of events in nerve nets and finite automata. *Automata studies,* Том 34, pp. 3--41.

Kocher, M. & Savoy, J., 2018. Distributed language representation for authorship attribution. *Digital Scholarship in the Humanities,* 33(2), pp. 425--441.

Kondratyuk, D. & Straka, M., 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099.*

Kornai, A., 1985. Natural languages and the Chomsky hierarchy. У: *Second Conference of the European Chapter of the Association for Computational Linguistics.* s.l.:s.n., pp. 1--7.

Krauwer, S. & Hinrichs, E., 2014. The CLARIN research infrastructure: resources and tools for e-humanities scholars. У: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).* s.l.:s.n., pp. 1525--1531.

Krstev, C., 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* s.l.:Faculty of Philology of the University of Belgrade.

Krstev, C., 2021. The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata. *Infotheca - Journal for Digital Humanities,* 21(2), pp. 26--42.

Krstev, C. & Stanković, R., 2020. Old or new, we repair, adjust and alter (texts). *Infotheca - Journal for Digital Humanities,* 12(2), pp. 61--80.

Krstev, C. & Stanković, R., 2022. *Report on the Serbian Language,* s.l.: European Language Equality (ELE).

Krstev, C. & Vitas, D., 2005. Corpus and lexicon-mutual incompletness. *Proceedings of the Corpus Linguistics Conference*, Том 14, p. 17.

Krstev, C., Vitas, D. & Erjavec, T., 2004. Morpho-Syntactic Descriptions in MULTEXT-East-the Case of Serbian.. *Informatica (Slovenia),* 28(4), pp. 431--436.

Krstev, C., Vitas, D. & Trtovac, A., 2011. Orwell's 1984–the Case of Serbian Revisited. У: *Proceedings of 5th Language & Technology Conference.* s.l.:s.n.

Kuhn, R. & De Mori, R., 1990. A cache-based natural language model for speech recognition. *IEEE transactions on pattern analysis and machine intelligence,* 12(6), pp. 570--583.

Langner, B., 2010. *Data-driven Natural Language Generation: Making Machines Talk Like Humans Using Natural Corpora.* s.l.:Language Technologies Institute, Carnegie Mellon University.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *nature,* 521(7553), pp. 436--444.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. У: *Text summarization branches out.* s.l.:s.n., pp. 74--81.

Liu, J. и други, 2015. Mining Quality Phrases from Massive Text Corpora. *SIGMOD '15 Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1729-1744.

Liu, X., He, P., Chen, W. & Gao, J., 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint 1901.11504.*

Liu, Y. и други, 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Lu, J., Batra, D., Parikh, D. & Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems,* Том 32.

Lyons, J., 1981. *Language and Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Manning, C. D. и други, 2014. The Stanford CoreNLP natural language processing toolkit. У: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.* s.l.:s.n., pp. 55--60.

Manning, C. & Schutze, H., 1999. *Foundations of statistical natural language processing.* s.l.:MIT press.

Marcus, S., 1997. Contextual grammars and natural languages. У: *Handbook of formal languages.* s.l.:Springer, pp. 215--235.

McCulloch, W. S. & Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics,* 5(4), pp. 115--133.

Mendenhall, T. C., 1887. The characteristic curves of composition. *Science,* Issue 214c, pp. 237--246.

Milovanović, B. & Stanković, a. R., 2020. Part of Speech Tagging for Serbian language using Natural Language Toolkit. *History,* Том 5.

Mizumoto, M., Toyoda, J. & Tanaka, K., 1972. *Information sciences,* 4(1), pp. 87--100.

Moretti, F., 2000. Conjectures on world literature. *New left review,* 1(54).

Mykowiecka, A., 1991. Natural-language generation - an overview. *Int. J. Man-Machine Studies,* Том 34, pp. 497-511.

O'Shea, K. & Nash, R., 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458.*

Paperno, D. и други, 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031.*

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. У: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* s.l.:s.n., pp. 311--318.

Park, J. & Sandberg, I. W., 1991. Universal Approximation Using Radial-Basis-Function. *Neural computation,* 3(2), pp. 246--257.

Paumier, S., Nakamura, T. & Voyatzi, S., 2009. Unitex, a corpus processing system with multi-lingual linguistic resources. *eLEX2009,* Том 173.

Perišić, O. и други, 2022. *It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian.* s.l.:s.n.

Pitler, E. & Nenkova, A., 2008. Revisiting readability: a unified framework for predicting text quality. *08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 186-195.

Qi, P. и други, 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082.*

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., 2018. *Improving language understanding by generative pre-training.* s.l.:OpenAI.

Radford, A. и други, 2019. Language models are unsupervised multitask learners. *OpenAI blog,* 1(8).

Rybicki, J. & Eder, M., 2011. Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and linguistic computing,* pp. 315--321.

Sakaguchi, K., Bras, R. L., Bhagavatula, C. & Choi, Y., 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM,* 64(9), pp. 99--106.

Salami, D. & Momtazi, S., 2021. Recurrent convolutional neural networks for poet identification. *Digital Scholarship in the Humanities,* 36(2), pp. 472--481.

Šandrih, B., Krstev, C. & Stanković, R., 2019. Development and evaluation of three named entity recognition systems for serbian-the case of personal names. У: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019).* s.l.:s.n., pp. 1060--1068.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Schmid, H., 1999. Improvements in part-of-speech tagging with an application to German. У: *Natural language processing using very large corpora.* s.l.:Springer, pp. 13--25.

Schmid, H., 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. У: *Proceedings of the 3rd international conference on digital access to textual cultural heritage.* s.l.:s.n., pp. 133-137.

Schöch, C., Erjavec, T., Patras, R. & Santos, D., 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open.*

Segarra, S., Eisen, M. & Ribeiro, A., 2015. Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing,* 63(20), pp. 5464--5478.

Sellam, T., Das, D. & Parikh, A., 2020. BLEURT: Learning Robust Metrics for Text Generation. У: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* s.l.:s.n., pp. 7881--7892.

Shen, D., 2022. FoundationLayerNorm: Scaling BERT and GPT to 1,000 Layers. *arXiv preprint arXiv:2204.04477.*

Sinclair, J., 1984. Naturalness in language. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies,* 5(11), pp. 45-55.

Sinclair, J., 1991. *Corpus Concordance and Collocation.* s.l.:Oxford University Press.

Sjöbergh, J., 2003. Combining POS-taggers for improved accuracy on Swedish text. У: *Proceedings of NODALIDA.* s.l.:s.n.

Škorić, M. и други, 2022. Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics,* 10(5), p. 838.

Song, K. и други, 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450.*

Stamatatos, E., 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology,* 60(3), pp. 538--556.

Stamatatos, E. и други, 2014. Overview of the author identification task at PAN 2014. У: *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014.* s.l.:s.n., pp. 1--21.

Stanković, R., Krstev, C., Lazić, B. & Škorić, M., 2018. Electronic Dictionaries - from File System to lemon Based Lexical Database. *Proceedings of the 11th International Conference on Language Resources and Evaluation - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018)*, pp. 48--56.

Stanković, R., Krstev, C., Šandrih Todorović, B. & Škorić, M., 2021. Annotation of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities,* 21(2), pp. 43--59.

Stanković, R. и други, 2020. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. У: *Proceedings of The 12th Language Resources and Evaluation Conference.* s.l.:s.n., pp. 3954--3962.

Stanković, R., Škorić, M. & Šandrih Todorović, B., 2022. Parallel Bidirectionally Pretrained Taggers as Feature Generators. *Applied Sciences,* 12(10), p. 5028.

Sun, C. и други, 2019. Videobert: A joint model for video and language representation learning. У: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* s.l.:s.n., pp. 7464--7473.

Sundermeyer, M., Schlüter, R. & Hermann, N., 2012. LSTM neural networks for language modeling. У: *Thirteenth annual conference of the international speech communication association.* s.l.:s.n.

Sun, Y. и други, 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint 1904.09223.*

Teofili, T., 2019. Document embeddings for rankings and recommendations. У: *Deep learning for search.* s.l.:Simon and Schuster, p. 6.1.

Tsoulos, I., Gavrilis, D. & Glavas, E., 2008. Neural network construction and training using grammatical evolution. *Neurocomputing,* 72(1-3), pp. 269-277.

Turing, A., 1938. On computable numbers, with an application to the Entscheidungsproblem. A correction. *Proceedings of the London Mathematical Society,* 2(1), pp. 544--546.

Turing, A. M., 1950. Computing Machinery and Intelligence. *Mind,* Том 49, pp. 433-460.

Utvić, M., 2011. Annotating the Corpus of Contemporary Serbian. У: *Proceedings of the INFOtheca '12 Conference.* s.l.:s.n., pp. 36--47.

Utvić, M., 2014. *Izgradnja referentnog korpusa savremenog srpskog jezika,* s.l.: Универзитет у Београду.

Vaswani, A. и други, 2017. *Attention Is All You Need.* s.l.:arXiv.

Vitas, D. & Krstev, C., 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne,* Том 63, p. 279–292.

Vogler, C. & Metaxas, D., 1999. *Parallel hidden Markov models for American sign language recognition.* New York, IEEE, pp. 116-122.

Wang, B. & Komatsuzaki, A., 2021. *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.* s.l.:s.n.

Watrous, R. L. & Kuhn, G. M., 1992. Induction of Finite-State Languages Using Second-Order. *Neural Computation,* Том 4, pp. 406-414.

Weerasinghe, J. & Greenstadt, R., 2020. Feature vector difference based neural network and logistic regression models for authorship verification. У: *CEUR workshop proceedings*. s.l.:s.n.

Wexler, K. & Culicover, P. W., 1980. *Formal principles of Language Acquisition*. Cambridge: MIT Press.

Wolf, T. и други, 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, Z. и други, 2019. understanding, Xlnet: Generalized autoregressive pretraining for language. *Advances in neural information processing systems*, Том 32.

Zellers, R. и други, 2019. HellaSwag: Can a machine really finish your sentence?. *arXiv preprint arXiv:1905.07830*.

Zhang, T. и други, 2019. BERTScore: Evaluating Text Generation with BERT. У: *International Conference on Learning Representations*. s.l.:s.n.

Zhang, Y., 2004. *Using Bayesian Priors to Combine Classifiers for Adaptive Filtering*. New York, Association for Computing Machinery, p. 345–352.

Zhang, Z. и други, 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Zhu, Y. и други, 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. У: *Proceedings of the IEEE international conference on computer vision*. s.l.:s.n., pp. 19--27.

Левенштейн, В. И., 1965. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академии наук*, 163(4), pp. 845--848.

Утвић, М., 2013. Листе учестаности Корпуса савременог српског језика. *Научни састанак слависта у Вукове дане. Српски језик и његови ресурси: теорија, опис и примене*, 43(3), p. 241–262.

БИОГРАФИЈА АУТОРА:

Михаило Шкорић рођен је 1992. године у Београду. Одрастао је у Обреновцу, где је похађао гимназију и матурирао 2011. године. Исте године уписује основне академске студије на Филолошком факултету Универзитета у Београду, модул Библиотекарство и Информатика, на смеру Језик, књижевност, култура. Дипломирао је 2015. године са просечном оценом 8.65 и уписао мастер академске студије истог профила. Завршни рад под називом „Сврставање појмова на скали позитивно-негативно, применом истраживања података над корпусом текстова који садрже емотиконе“ одбранио је 2016. године, и завршио мастер академске студије са просечном оценом 10. Исте године уписује интердисциплинарне докторске студије Универзитета у Београду, смер Интелигентни системи.

Од 2017. године ради у Рачунарском центру Рударског одсека на Рударско-геолошком факултету, у сфери системског и софтверског инжењерства, а исте године стиче и звање истраживач приправник при катедри за примењену математику и информатику. Звање истраживач-сарадник стиче 2020. године, када пријављује тему докторске дисертације „Композитне псеудограматике засноване на паралелним језичким моделима српског језика“. Аутор је и коаутор осамнаест научних радова из области обраде природног језика, интелигентних система и библиотечко-информационих наука, објављених у домаћим и страним часописима и монографијама, од којих су 3 са СЦИ листе.

У оквиру COST акције IC1302 – *Keystone*, учествује на летњој школи *Keyword search in big linked data* на Техничком Универзитету у Бечу. Био је активан учесник COST акције CA16204 – *Distant Reading for European Literary History*, у оквиру које учествује на четвородневном тренингу метода и техника удаљеног читања на Националном Универзитету Ирске у Галвеју 2018, као и на краткотрајној научној мисији на Институту за Пољски језик у Кракову у марту 2020. године, са темом компаративне стилистичке и морфосинтаксне анализе текстова. Активан је члан Друштва за језичке ресурсе и технологије – JePTex – где учествује у развоју система и алата за обраду српског језика.

образац изјаве о ауторству

Изјава о ауторству

Име и презиме аутора Михаило Шкорић

Број индекса 42/2016

Изјављујем

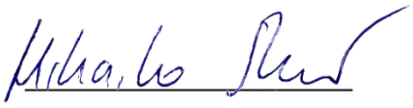
да је докторска дисертација под насловом

Композитне псеудограматике засноване на паралелним језичким моделима српског језика

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 20.1.2023.



Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Михаило Шкорић

Број индекса 42/2016

Студијски програм Интелигентни системи

Наслов рада Композитне псеудограматике засноване на паралелним језичким моделима српског језика

Ментор Проф. др. Ранка Станковић

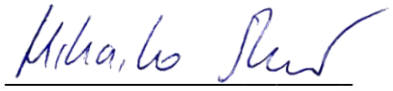
Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 20.1.2023.



образац изјаве о коришћењу

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Композитне псеудограматике засноване на паралелним језичким моделима српског језика

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

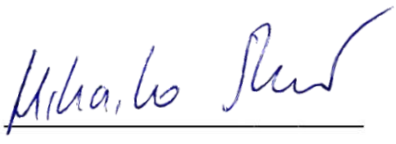
5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 20.1.2023.



1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.