

УНИВЕРЗИТЕТ У БЕОГРАДУ

Саша З. Петалинкар

Машинско учење и велики језички модели у
развоју семантичких мрежа и њиховој
примени на аутоматско разумевање текста

Докторска дисертација

Београд, 2026.

UNIVERSITY OF BELGRADE

Saša Z. Petalinkar

Machine Learning and Large Language
Models in the Development of Semantic
Networks and the Application of These
Networks to Automatic Text Understanding

Doctoral Dissertation

Belgrade, 2026

Ментори:

др Ранка Станковић, редовни професор

Универзитет у Београду, Рударско-геолошки факултет

др Јелена Граовац, ванредни професор

Универзитет у Београду, Математички факултет

Чланови комисије:

др Владан Девеџић, академик, редовни професор

Универзитет у Београду, Факултет организационих наука

др Ненад Митић, редовни професор у пензији

Универзитет у Београду, Математички факултет

др Александра Марковић, виши научни сарадник

Институт за српски језик САНУ

Датум одбране: _____

Захвалница

Захваљујем својим менторкама, проф. др Ранки Станковић и проф. др Јелени Граовац, на поверењу, стручном усмеравању и подршци током израде ове докторске дисертације. Њихови савети и примедбе помогли су ми да прецизније поставим истраживачка питања, уобличим методологију и јасније изложим резултате. Посебну захвалност дугујем проф. др Ранки Станковић за подстицај да истраживање повежем са развојем језичких ресурса за српски језик, за научну сарадњу на радовима из којих је проистекао део резултата дисертације и за истрајну подршку у различитим фазама рада. Проф. др Јелени Граовац захваљујем на стручним саветима и пажњи посвећеној уобличавању дисертације.

Проф. др Цветани Крстев захваљујем на помоћи у раду са Српским ворднетом, на стручним сугестијама и на сарадњи у истраживању разрешења вишезначности речи. Њено познавање српског језика и искуство у изградњи лексичких ресурса били су драгоцен ослонац при тумачењу и провери језичке грађе.

Захваљујем члановима комисије, проф. др Владану Девеџићу, проф. др Ненаду Митићу и др Александри Марковић, на времену и пажњи посвећеним прегледу и оцени ове докторске дисертације.

Милици Иконић Нешић захваљујем на научној сарадњи у истраживањима лексикона сентимента и разрешења вишезначности речи, као и на учешћу у ручној евалуацији и стручној провери података. Александри Марковић захваљујем на времену, знању и труду уложеним у анотацију и проверу језичке грађе. Захвалност упућујем и осталим сарадницима који су учествовали у припреми и провери корпуса ELEXIS-WSD-sr и пратећих лексичко-семантичких ресурса.

Посебну захвалност дугујем студенткињама и студентима Филолошког факултета Универзитета у Београду за учешће у ручном означавању сентимента новинских реченица и у евалуацији синтетичког скупа синсетова. Њихов труд, пажљиве процене и одговоран однос према задацима били су драгоцени у провери језичке грађе коришћене у овом истраживању.

Михаилу Шкорићу захваљујем на научној сарадњи на заједничким радовима о текстуалним корпусима и језичким моделима. Проф. др Ивану Обрадовићу захваљујем на помоћи при ревизији рада о лексиконима сентимента, чији су резултати обухваћени дисертацијом. Захвалност дугујем и сарадницима Друштва за језичке ресурсе и технологије — JePTex за развој ресурса за српски језик и њихову доступност за ова истраживања.

Део истраживања представљених у овој дисертацији реализован је уз подршку Фонда за науку Републике Србије, у оквиру пројекта „Text Embeddings – Serbian Language Applications“ (TESLA, број 7276). Захваљујем и Рударско-геолошком факултету Универзитета у Београду на подршци истраживању разрешења вишезначности речи, чији су резултати обухваћени овом дисертацијом.

Својој породици захваљујем на љубави, стрпљењу и разумевању током докторских студија и израде дисертације. Њихова подршка била ми је ослонац у периодима неизвесности и охрабрење да истрајем до завршетка рада. Захвалан сам и пријатељима који су ме бодрили и имали разумевања за време и пажњу које је овај рад захтевао.

ПОДАЦИ О ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

Наслов докторске дисертације: Машинско учење и велики језички модели у развоју семантичких мрежа и њиховој примени на аутоматско разумевање текста

Сажетак:

У дисертацији се испитује примена метода машинског учења и великих језичких модела у развоју, ревизији и примени лексичко-семантичких мрежа, са фокусом на Српски ворднет. Полазиште рада је потреба да се ресурси за српски језик унапреде не само повећањем обима, већ и прецизнијим моделовањем значења, сентимента и односа између лексичких јединица. Истраживање обухвата изградњу и евалуацију слоја сентимента Српског ворднета, формирање синтетичког скупа за евалуацију сентимента, развој инвентара значења и поступака за разрешење вишезначности речи, дестилацију знања у мање моделе, превод и допуну синсета, као и поравнање Српског ворднета са описним речником. Посебно се испитује у којој мери се обogaћени ресурс може применити у конкретним задацима који користе његове резултате, односно у одређивању сентимента израженог у тексту и аутоматском превођењу текста са једног језика на други. Резултати показују да комбинација експлицитног лексичко-семантичког знања, ручно проверених евалуационих скупова и контролисане употребе великих језичких модела може да подржи развој Српског ворднета и аутоматско обележавање текста синсетима. Истовремено, резултати указују да потпуно аутоматско проширивање ресурса није довољно поуздано без лексикографске ревизије, јасних критеријума прихватања и додатне емпиријске провере у различитим доменима. Дисертација зато доприноси и развоју конкретних ресурса за српски језик и методолошком разумевању начина на који се симболичко знање и савремени језички модели могу комбиновати у задацима аутоматског разумевања текста.

Кључне речи: Српски ворднет; семантичке мреже; машинско учење; велики језички модели; разрешење вишезначности речи; анализа сентимента; SentiWordNet; лексичко-семантичко поравнање; обрада природног језика; српски језик.

Научна област: Интелигентни системи

Ужа научна област: Обрада природног језика, Рачунарска лингвистика

DISSERTATION INFORMATION

Dissertation title: Machine Learning and Large Language Models in the Development of Semantic Networks and the Application of These Networks to Automatic Text Understanding

Abstract:

This dissertation investigates the use of machine learning methods and large language models in the development, revision, and application of lexical-semantic networks, with a focus on Serbian WordNet. The starting point of the research was the need to improve resources for Serbian not only by increasing their size, but also by modelling word meanings, sentiment, and lexical-semantic relations more precisely. The research covers the construction and evaluation of a sentiment layer for Serbian WordNet, the creation of a synthetic sentiment evaluation set, the development of a sense repository and procedures for word sense disambiguation, knowledge distillation into smaller models, synset translation and enrichment, and the alignment of Serbian WordNet with a descriptive dictionary. Special attention is given to the extent to which the enriched resource can be used in downstream tasks such as sentiment analysis and machine translation. The results show that combining explicit lexical-semantic knowledge, manually checked evaluation sets, and controlled use of large language models can support the development of Serbian WordNet and the automatic annotation of text with synsets. At the same time, the findings indicate that fully automatic resource expansion is not sufficiently reliable without lexicographic revision, clear acceptance criteria, and further empirical validation across domains. The dissertation therefore contributes both to the development of concrete resources for Serbian and to the methodological understanding of how symbolic knowledge and contemporary language models can be combined in automatic text understanding.

Key words: Serbian WordNet; semantic networks; machine learning; large language models; word sense disambiguation; sentiment analysis; SentiWordNet; lexical-semantic alignment; natural language processing; Serbian.

Scientific field: Intelligent Systems

Scientific subfield: Natural Language Processing, Computational Linguistics

Садржај

1	Увод	1
2	Преглед досадашњих истраживања	5
2.1	Обрада природног језика као истраживачка и примењена област	5
2.2	Значење и вишезначност у природном језику	6
2.3	Лексичко-семантички ресурси и Принстонски ворднет.....	8
2.4	Ворднети других језика и вишејезичке семантичке мреже	10
2.5	SentiWordNet и лексичко-семантичка репрезентација сентимента	11
2.6	Методе изградње SentiWordNet-a	13
2.7	Савремени приступи машинског учења и велики језички модели.....	16
2.8	Фино подешавање и закључивање	19
2.9	Обликовање промптова моделима	20
2.10	Генерисање података и дестилација модела	23
2.11	Ресурси и приступи за разрешење вишезначности речи	24
2.12	Српски ворднет: стање, значај и изазови	28
2.13	Аутоматско и полуаутоматско проширивање ворднета у другим језицима	29
2.14	Поравнање лексичко-семантичких ресурса.....	30
3	Методологија	34
3.1	Оцена сентимента у ворднету	34
3.2	Скуп за евалуацију оцена сентимента креиран генеративним моделом.....	40
3.3	Разрешење вишезначности речи – генеративни модел	44
3.4	Разрешење вишезначности речи – дестилација	49
3.5	Превод синсета	51
3.6	Поравнање лексичко-семантичких ресурса уз употребу генеративних модела.....	54
3.7	Екстерна евалуација обогаћеног Српског ворднета у анализи сентимента	55
3.8	Циљана провера употребе слоја за разрешење вишезначности речи у машинском превођењу	58
4	Коришћени скупови података и велики језички модели	60
4.1	Принстонски ворднет	61
4.2	Српски ворднет	61
4.3	ELEXIS-WSD и ELEXIS-WSD-sr	62
4.4	Речник савременог српског језика	64
4.5	Senti-pol-sr.....	65
4.6	Коришћени велики језички модели	66
4.7	Скуп новинских реченица за евалуацију сентимента.....	68
5	Резултати	70
5.1	Омотач за ворднет	70

5.2	Евалуација модела обучених за процену сентимента.....	71
5.2.1	Евалуација добијених лексикона сентимента	73
5.2.2	Ручна евалуација добијених лексикона сентимента	73
5.2.3	Формирање и провера скупа за евалуацију сентимента синсета	74
5.2.4	Евалуација добијених лексикона на синтетичком евалуационом скупу.....	78
5.3	Изградња почетног инвентара значења WSD-sr-sens-inv-v1	80
5.4	Унапређење инвентара и његова улога у анотацији	83
5.5	Опис златног скупа и анализа инвентара значења	85
5.6	Квантитативна евалуација перформанси	90
5.7	Квалитативна анализа грешака и анализа вишезначних лема.....	101
5.8	Резултати дестилације модела за разрешење вишезначности речи	103
5.9	Резултати формирања српских синсета помоћу великих језичких модела	108
5.10	Резултати поравнања лексичко-семантичких ресурса са описним речником	113
5.11	Екстерна евалуација обогаћеног Српског ворднета у анализи сентимента	119
5.12	Циљана провера употребе слоја за разрешење вишезначности речи у машинском превођењу	122
6	Дискусија	127
6.1	Лексикони сентимента.....	127
6.2	Разрешење вишезначности речи и редакција инвентара значења	131
6.3	Дестилација модела и ограничења мањих модела	136
6.4	Превод синсета помоћу великих језичких модела.....	139
6.5	Поравнање лексичко-семантичких ресурса.....	141
6.6	Низводни задаци и границе употребе обогаћеног ресурса	145
7	Закључак	148
7.1	Доприноси	150
7.2	Ограничења и даљи рад.....	151
8	Литература	154
9	Прилог термилошки глосар	166
10	Прилог индекс табела	179
11	Прилог индекс слика.....	181

1 Увод

Аутоматска обрада природног језика (ОПЈ; енгл. *natural language processing*, NLP) једно је од најдинамичнијих поља савремене информатике, са дубоким утицајем на науку, привреду и друштво у целини. Велики језички модели (ВЈМ; енгл. *large language models*, LLM) постали су инфраструктурна окосница широког спектра примена, укључујући машинско превођење и сажимање текста (сумаризацију), извлачење знања и класификацију текста, као и анализу сентимента и разговорне системе. Иако ВЈМ могу да производе граматички исправан и семантички уверљив текст, њихова стабилност, доменска преносивост и објашњивост и даље зависе од експлицитно моделованог лексичко-семантичког знања. Ту у први план ступају уређени лексички и семантички ресурси, попут семантичких мрежа типа ворднет (енгл. *WordNet*) (Miller 1995), који дефинишу значења (синсете) и односе између њих (синонимија, хиперонимија/хипонимија, меронимија и др.). За глобалне језике, пре свега енглески, развијени су обимни и систематично семантички структурирани ресурси од којих је најпознатији Принстонски ворднет (енгл. *Princeton WordNet*, скр. PWN) (Fellbaum 1998), што омогућава напреднију интеграцију симболичког знања у статистичке моделе.

EuroWordNet (Fellbaum 2005; Vossen 2004) је поставио основу за вишејезичке језичке семантичке мреже са јединственим идентификатором концепта у свим језицима, међујезичким индексом (енгл. *Inter-Lingual Index*, скр. ILI) који је уједно као стожер усвојен у свим каснијим пројектима. Српски ворднет (SrpWN) (Krstev et al. 2004) креиран је у оквиру пројекта BalkaNet (Tufis et al. 2004) који је усвојио шему пројекта EuroWordNet. Верзија мреже Српски ворднет из 2018. године је садржала 25,320 синсета, међу којима се, осим синсета мапираних са PWN, такође налазе и они са значењима карактеристичним за језике обухваћене пројектом BalkaNet (Krstev et al. 2004; Stanković, Mladenović, et al. 2018). На основу повезаности Српског ворднета са Принстонским ворднетом, вредности сентимента¹ из ресурса SentiWordNet (SWN) (Baccianella et al. 2010) пресликане су на одговарајуће синсете Српског ворднета, чиме је овај ресурс обogaћен информацијама о сентименту. Сваки синсет у мрежи Српски ворднет се проширује вредношћу сентимента, следећи поступак предложен у (B. Liu 2010). Пошто је SentiWordNet базиран на мрежи PWN, синсети који су карактеристични за BalkaNet, као и они који су карактеристични само за српски и нису поравнати са PWN тако да им нису додељене вредности сентимента. Вредности сентимента из мреже Српски ворднет коришћене су за анализу сентимента за текстове на српском језику (Mladenović et al. 2015).

Иако Српски ворднет представља значајан ресурс у обради природног језика и рачунарској лингвистици и користи се за решавање различитих задатака за српски језик, његову примену и даље ограничава његова скромна величина, јер је око четири пута мањи од мреже PWN. Проширење ове семантичке мреже није једноставан задатак и захтева значајне људске ресурсе, јер спада у ред ресурса који се ручно развијају (Krstev et al. 2004). Било је и раније покушаја да се изградња Српског ворднета аутоматизује (Stanković, Mladenović, et al. 2018), али је предложени процес још увек захтеван, те је потреба за ефикаснијом аутоматизацијом неопходна. Проблем који до сада није решаван за српски језик је обележавање значења текстуалних речи (моноксемских или полилексемских јединица) уз разрешавање

¹ У анализи сентимента, сентимент означава вредносну усмереност става израженог у тексту према одређеном ентитету, теми или њиховом аспект и најчешће се одређује као позитиван, негативан или неутралан. Он се разликује од емоције, јер описује пре свега поларитет става, док емоција означава специфично афективно стање, попут радости, туге, страха или љутње.

вишезначности, што је једна од веома важних примена Српског ворднета у аутоматском разумевању текста.

Речи (јединице лексикона) могу имати једно или више значења (Mišić et al. 2024), а у ворднету је свако значење представљено посебним синсетом, па тако једној текстуалној речи може одговарати више синсета, али и један синсет може бити описан помоћу једне или више текстуалних речи², тј. блискозначница које описују конкретан појам. Према томе, неопходно је дефинисати поступак који свакој текстуалној речи придружује одговарајуће значење, односно синсет.

Развој и унапређење лексичких ресурса за српски језик, у које спада и Српски ворднет, стога представљају не само научни него и друштвени приоритет. Традиционални поступци изградње захтевају значајне људске и временске ресурсе, што ограничава обим и темпо изградње, па самим тим и могућност практичне примене. Стога ова докторска дисертација истражује могућности примене метода машинског учења и BJM (Devlin et al. 2019; Grattafiori et al. 2024; Škorić and Petalinkar 2024; Vreš et al. 2024; Škorić 2025) на аутоматизацију развоја и унапређења семантичке мреже Српски ворднет, као и за њихово коришћење у аутоматском разумевању текста. Посебан фокус је стављен на развој алата за аутоматску анотацију текста којим се текстуалним речима (моноксемским или полилексемским) придружују синсети Српског ворднета (као што је то предложено у (Orkphol and Yang 2019) за енглески језик), као и на истраживање могућности проширивања ворднета интеграцијом информација из мреже PWN и других лексичких извора (Sun and Wang 2025; Al-Edhari and Kahani 2025; Obare et al. 2025). Такође се предлаже и упаривање према значењу (енгл. *word sense alignment*) (Oliver 2020; Vij et al. 2025) мреже Српски ворднет са описним речницима српског језика. Резултати овог истраживања дају допринос не само лингвистичкој науци већ и применама у интелигентним системима, обради текста и образовању.

Циљ истраживања представљеног у овој дисертацији је развој методологије за аутоматизацију изградње и унапређења семантичких мрежа, као и за њихову примену, са фокусом на Српски ворднет. Ова методологија обухвата ревизију постојећих вредности сентимента, проширење синсета новим литеалима (речима, односно лексичким јединицама), дефиницијама и примерима, интеграцију спољних ресурса, као и препознавање значења текстуалних речи. Иако се емпиријска верификација спроводи на српском, предложени поступци не зависе од језика па се они могу применити и на друге језике, а посебно оне са слабије развијеним језичким ресурсима. Српски ворднет служи као главна студија случаја — као доказ изводљивости, корисности и ограничења — док се језички специфичне компоненте (лематизација, морфолошка анализа, правописне варијанте) систематски издвајају у опису појединачних метода. На тај начин рад истовремено унапређује аутоматско разумевање текста на српском и нуди применљив образац за одржив развој семантичких ресурса у широј заједници језика са недовољно развијеном лингвистичком инфраструктуром.

Специфични циљеви укључују:

- Развој аутоматског анотатора текста заснованог на BJM и методама машинског учења, који ће омогућити повезивање речи у тексту са одговарајућим значењем представљеним синсетом.

² Под лексичком јединицом у овом контексту подразумевамо улаз у речник, док је литеал, који је појам из ворднета, једно значење те речи (обично у описним речницима дато под 1, 2, итд. па затим а, б, итд.). Текстуална реч представља појављивање речи у тексту (може бити и у неком флективном облику), рецимо текстуални токен.

- Замену постојећих вредности сентимента, добијених проширењем Српског ворднета вредностима из SentiWordNet-a, вредностима добијеним машинским учењем.
- Примену савремених ВЈМ на задацима аутоматског усклађивања синсета са мрежом PWN.
- Испитивање могућности примене генеративних модела за допуну Српског ворднета.
- Тестирање примене обогаћеног Српског ворднета на задацима анализе сентимента, класификације текста и машинског превођења.
- Лексичко поравнавање Српског ворднета са описним речником српског језика.

Циљ је прављење ефикасног система који интегрише постојеће језичке моделе и који се може надоградити језичким моделима који ће се у будућности развијати.

Полазећи од наведених циљева и дефинисаног истраживачког оквира, формулисаћемо хипотезе које се у дисертацији емпиријски испитују. Хипотезе су организоване као једна општа (надређена) хипотеза и више помоћних хипотеза које је могуће директно операционализовати у експерименталном делу рада.

H0 (општа хипотеза). Применом метода машинског учења и великих језичких модела могуће је (i) у значајној мери аутоматизовати и унапредити кључне процесе развоја семантичких мрежа типа ворднет за српски језик и (ii) омогућити поуздано аутоматско обележавање текста чворовима (синсетима) мреже. Очекује се да ће обогаћена верзија Српског ворднета остварити боље перформансе у односу на полазну (енгл. *baseline*) верзију у задацима повезивања значења, претраживања и класификације текста.

H1 (сентимент). Оцене сентимента синсета добијене методом машинског учења (обућавањем/калибрацијом на ресурсима релевантним за српски језик) статистички значајно боље одражавају употребу и сентимент значења представљеног српским синсетом у српском језику него оцене преузете непосредним пресликавањем из SentiWordNet-a.

H2 (векторска репрезентација значења и поравнање значења речи). У контекстуалном векторском простору могуће је конструисати репрезентације које омогућавају одређивање значења циљне лексичке јединице, односно њено поуздано придруживање одговарајућем синсету Српског ворднета, како на нивоу текстуалних појава (енгл. *token-level*), тако и на нивоу значења у описном речнику (енгл. *sense-level*) у контексту лексичко-семантичког поравнавања.

H3 (генеративна допуна и усклађивање). Велики језички модели могу се употребити за контролисану допуну синсета (нпр. генерисање примера употребе, предлог синонимних/лексикалних варијанти, формулацију дефиниција), као и за усклађивање/мапирање значења из других речника са енглеским еквивалентима у PWN (где је могуће), при чему се очекује да ће такав приступ бити робуснији од директног машинског превода у случајевима лексичко-семантичких неподударања између језика.

H4 (низводни задаци (енгл. *downstream tasks*)). Обогаћена верзија Српског ворднета, добијена предложеним поступцима, доводи до мерљивог и статистички значајног побољшања резултата у одабраним низводним задацима ОПЈ за српски језик (нпр. анализа сентимента, класификација текста, откривање лажних вести, повезивање именованих ентитета са базом знања), у поређењу са резултатима добијеним употребом полазне верзије ресурса.

У оквиру ове дисертације хипотеза *H4* операционализује се кроз два одабрана низводна задатка³: анализу сентимента и машинско превођење текста. За потребе њихове емпиријске провере формулишу се две уже радне потхипотезе:

H4a. Анализа сентимента. У задатку анализе сентимента реченица, употреба коначне обогачене верзије Српског ворднета са вредностима сентимента на нивоу синсета даје боље резултате од приступа заснованог на вредностима сентимената придружених лемама, као и од приступа заснованог на полазним вредностима пресликаним из енглеског SentiWordNet-a.

H4b. Циљана провера у машинском превођењу. У задатку превођења са српског на енглески језик, експлицитна информација о значењу текстуалне речи преузета из инвентара значења може допринети правилнијем избору енглеског преводног еквивалента за лексички двосмислене јединице, у поређењу са преводом исте реченице без такве информације.

Научни доприноси ове дисертације обухватају развој и евалуацију слоја сентимента Српског ворднета, укључујући лексиконе и моделе за одређивање сентимента на нивоу синсета, као и утврђивање ограничења непосредног преношења оцена из SentiWordNet-a. Значајан допринос представљају и поступци и ресурси за разрешење вишезначности речи у српском језику: инвентари значења, ручно проверени златни скуп, сребрни подаци и дестиловани модели. Истраживање обухвата и методе превођења синсета и лексичко-семантичког поравнања описног речника са Српским ворднетом, чији су резултат скупови кандидата намењени експертској провери и даљој редакцији ресурса. Ови доприноси чине јединствен методолошки оквир за систематско проширивање, проверу и примену Српског ворднета у задацима обраде природног језика.

Структура текста дисертације организована је на следећи начин. Након уводног поглавља, у другом поглављу дат је преглед релевантних претходних истраживања и постојећих ресурса у области семантичких мрежа и аутоматског разумевања текста. Треће поглавље описује предложену методологију, укључујући коришћене моделе, ресурсе и поступке аутоматизације. У четвртом поглављу представљени су коришћени скупови података и начин њихове обраде. Пето поглавље приказује резултате експеримената, док су анализа и евалуација дате у наредном поглављу. У поглављу „Закључак” се сумирају доприноси рада и разматрају ограничења и могући правци даљих истраживања. Након закључка наведена је литература, а затим следе прилози: терминолошки глосар, индекс табела и индекс слика.

³ Низводни задаци (енгл. downstream tasks) јесу конкретни задаци обраде природног језика у којима се примењују претходно развијени модели, методе или језички ресурси, као што су анализа сентимента, машинско превођење, класификација текста и препознавање именованих ентитета. Њихови резултати могу послужити и за процену практичне употребљивости полазног модела или ресурса.

2 Преглед досадашњих истраживања

У овом поглављу разматрају се теоријски приступи, језички ресурси и методолошки оквири релевантни за истраживање представљено у овој дисертацији. Полазећи од ширег оквира обраде природног језика као истраживачке и примењене области, пажња се затим усмерава на питања значења и вишезначности, на лексичко-семантичке ресурсе и њихову улогу у формализацији знања, као и на савремене приступе засноване на машинском учењу и великим језичким моделима. Посебно место у овом прегледу имају приступи и ресурси релевантни за разрешење вишезначности, односно придруживања одговарајућег значења речи у контексту, као и анализа стања, значаја и ограничења Српског ворднета у ширем контексту развоја вишејезичких семантичких мрежа. Тако постављен преглед омогућава јасније позиционирање предмета и циљева овог истраживања у односу на постојећу литературу и доступне ресурсе.

2.1 Обрада природног језика као истраживачка и примењена област

Природни језик у дигиталном окружењу може се дефинисати као скуп изјава људи који су носиоци знања и намера, а које се у савременим информационим системима третирају као подаци за индексирање, претрагу, екстракцију информација и подршку доношењу одлука. Рад са овим дигитализованим подацима обједињује обрада природних језика: научна дисциплина и технологија која обухвата методе и алгоритме за аутоматско претварање текстуалних и говорних података у формалне репрезентације које машине могу да анализирају, интерпретирају и користе за извођење сложених задатака као што су класификација, означавање именованих ентитета, сажимање текста, машинско превођење, вођење дијалога и других облика комуникације између човека и машине (Jurafsky and Martin 2026).

ОПЈ данас представља фундаменталну основу рачунарских наука и вештачке интелигенције, јер њени темељи интегришу рачунарске методе (скалабилни алгоритми, машинско и дубоко учење), формалне лингвистичке дисциплине (морфологија, синтакса, семантика, прагматика) и приступе из когнитивних и статистичких наука (предикативни модели, представљања знања). Ова интердисциплинарност не само да води развоју алгоритама који обрађују језик, већ омогућава и практичну примену у широком спектру домена: од претраживања и индексирања корпуса до примена у областима од критичног значаја у здравству, праву, финансијама и јавном сектору.

Значај ОПЈ-а огледа се и у обиму и динамици научне продукције у овој области. Према подацима дигиталне библиотеке *ACL Anthology* (Association for Computational Linguistics 2026), која представља водећи репозиторијум радова из рачунарске лингвистике и обраде природног језика, база обухвата више од 120.000 научних публикација објављених на релевантним конференцијама, у часописима и у виду радних материјала. Континуирано повећање броја радова, укључујући хиљаде публикација годишње на водећим скуповима као што су ACL⁴, EMNLP⁵ и NAACL⁶, указује на стабилан раст истраживачког интереса и на институционализацију поља као једне од централних области савремених рачунарских наука.

Иако глобални развој ОПЈ-а показује високу динамику и технолошку зрелост, подршка језицима са мањим бројем говорника, укључујући српски, и даље је условљена ограниченом

⁴ Association for Computational Linguistics <https://www.aclweb.org/>

⁵ Conference on Empirical Methods in Natural Language Processing, <https://2026.emnlp.org/>

⁶ The Nations of the Americas Chapter of the Association for Computational Linguistics, <https://naacl.org/>

доступношћу дигиталних ресурса и инфраструктуре. Према извештају *European Language Equality* (Krstev and Stanković 2022) српски језик поседује одређени број развијених корпуса и алата за основну лингвистичку обраду, али у појединим категоријама, посебно у домену сложенијих семантичких и концептуалних ресурса, подршка је фрагментарна или ограничена. Ова диспропорција између глобалног напретка и локалне инфраструктуре указује на потребу за систематским развојем језичких ресурса као предуслова за стабилан напредак националних система за ОПЈ.

Практичне импликације ОПЈ-а у реалним системима су очигледне: модели за аутоматско извлачење чињеница убрзавају претрагу и анализу великих корпуса података, системи за одговарање на питање и вођење дијалога омогућавају интерактивну комуникацију са крајњим корисницима, а алати за резимирање, класификацију и детекцију намера помажу у управљању информационим преоптерећењем у организацијама и јавним службама. Такви системи омогућавају трансформацију података у знање, које представља суштински ресурс у дигиталном друштву (Young et al. 2018).

У исто време, експанзија истраживања открива и сложене изазове. Прво, високи резултати модела на стандардним тестовима сами по себи нису довољни да се утврди у којој мери модели поуздано тумаче значење исказа у контексту. Посебну проверу захтевају случајеви у којима тумачење зависи од намере говорника и знања о свету. Друго, контекстуализација и прагматичка тумачења (нпр. фигуративна, иронична или подразумевана значења) за аутоматске системе и даље представљају тешко решиве проблеме. Треће, зависност од корпуса за обучавање уводи и појачава пристрасности које могу имати негативне последице у примени. Коначно, показује се да је преносивост модела и њихова робусност у хетерогеним окружењима и окружењима са неконзистентним подацима ограничена у стварним индустријским и друштвеним сценаријима.

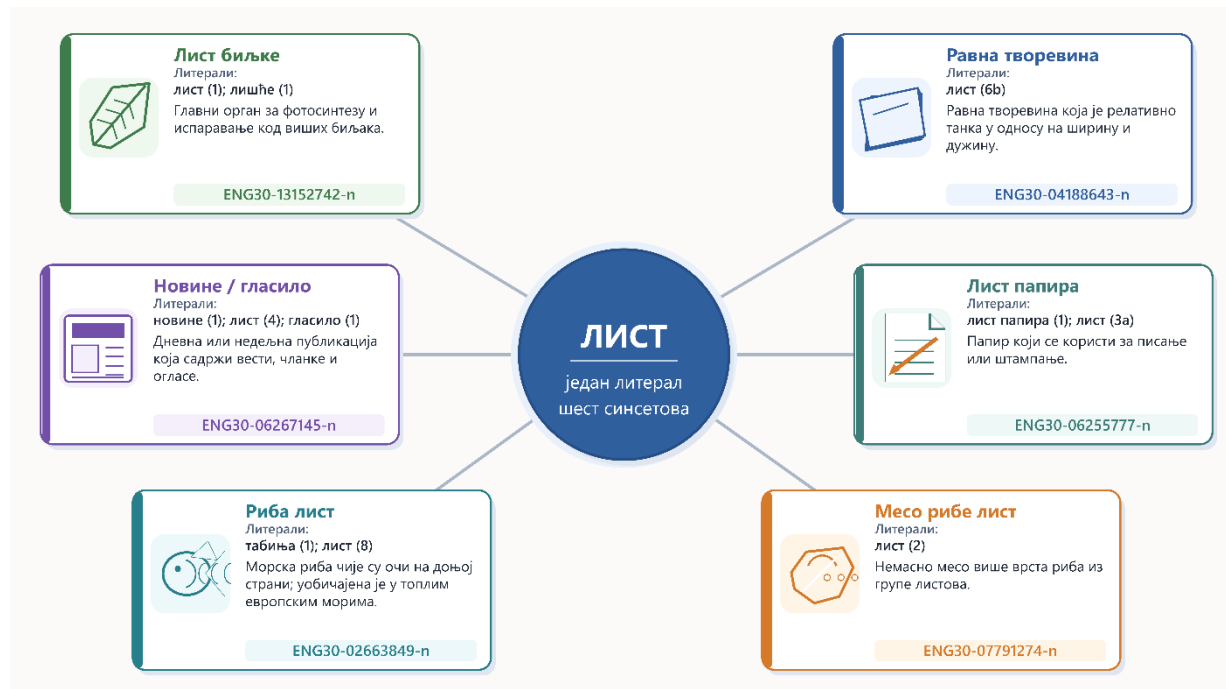
Упркос значајном технолошком напретку и експанзији истраживања, структурни изазови обраде природног језика и даље су у великој мери повезани са формалним представљањем значења. Високе перформансе статистичких модела не гарантују стабилну интерпретацију лексичко-семантичких односа, нити обезбеђују транспарентну структуру знања која би омогућила поуздано проширивање и адаптацију система у различитим језичким окружењима. Управо у домену систематске репрезентације значења концентрише се један од кључних структурних изазова савремене ОПЈ-а, што представља природну полазну тачку за даљу анализу модела репрезентације знања.

2.2 Значење и вишезначност у природном језику

Једна од основних карактеристика лексичког система природног језика јесте вишезначност, односно полисемија. У лексичкој семантици полисемија се обично дефинише као појава да једна лексема може бити повезана са два или више међусобно сродних значења (Cruse 2001, 80). Ова појава није изузетак већ системска особина језика, настала као резултат историјског развоја значења, метафоричких и метонимијских проширења, као и економичности лексичког фонда. У српској лингвистичкој традицији полисемија се такође описује као својство лексеме да обједињује више семантички повезаних значења која се у конкретној комуникацији активирају у зависности од контекста (Piper et al. 2022).

Слика 2.1 приказује један пример полисемије у српском језику може се представити на примеру именице *лист*. У зависности од контекста, ова реч може означавати лист биљке („лист

је пожутео у јесен“), лист папира („написао је одговор на једном листу папира“), као и новине или штампано гласило („прочитао је текст у дневном листу“) или рибу („риба лист се добро камуфлира“). Иако се ова значења односе на различите појмове, она су међусобно повезана семантичким развојем и метафоричким проширењима значења. Овакви примери показују да се различита значења једне речи не могу правилно интерпретирати без узимања у обзир ширег контекста у коме се појављују.



Слика 2.1 Шест значења лексеме „лист“ у Српском ворднету⁷.

У лингвистици се традиционално разликују полисемија, где су различита значења једне лексеме међусобно повезана, и хомонимија, где исти лексички облик има потпуно неповезана значења. На пример атлас:1 збирка географских карата (преузето из грчког), атлас:2 свилена тканина (преузето из турског). Међутим, у области обраде природног језика ова разлика се често не третира строго, јер оба феномена доводе до истог практичног проблема: потребе да се на основу контекста утврди које је значење речи актуелно у датом тексту. Хомографија, се односи на појаву да једна лексема има више међусобно повезаних значења, тако да се различите лексеме подударају у писаном облику, иако се могу разликовати по значењу, граматичким својствима и изговору. У речницима такве јединице стога се не представљају као различита значења исте одреднице, већ као засебне одреднице. Дobar пример представља *пут*. У значењу везаном за људско тело, као извор чулних осећања реч се изговара *пѹт* и именица је женског рода, док је *пѹт* у значењу друм, цеста именица мушког рода. Поред њих, хомографија је могућа и у другим врстама речи. У тексту без обележених акцената ове разлике нису видљиве: систем у свим случајевима добија исти низ карактера – *пут*. Због тога су, са становишта аутоматске обраде, полисемија и хомографија тесно повезани проблеми: у оба случаја потребно је на основу контекста утврдити којој лексичкој јединици, односно којем значењу, употребљени облик припада. Разрешавање хомографије при томе може захтевати не само семантичке већ и морфолошке и синтаксичке информације. Стога се у овом раду термин

⁷ Извор: Српски ворднет (<https://wn.jerteh.rs/>), претрага тачног литерала „list“, 23. 8. 2026. Синсет за лично име List (Франц Лист) није обухваћен.

полисемија користи у ширем смислу лексичке двосмислености, у складу са приступом који је уобичајен у истраживањима разрешења вишезначности речи (Navigli, 2009).

Полисемија има посебан значај у контексту аутоматске обраде текста. Један од првих истраживача који је указао на овај проблем у области рачунарске обраде језика био је Warren Weaver, који је у свом меморандуму о машинском превођењу из 1949. године истакао да исти лексички облик може имати различита значења у зависности од контекста, што представља један од основних изазова машинског превођења (Weaver 1949). Овај увид поставио је темеље каснијим истраживањима у рачунаској лингвистици, јер је показано да без контекстуалног одређивања значења није могуће постићи поуздану семантичку интерпретацију текста. Управо из овог проблема развио се задатак разрешења вишезначности речи (енгл. *Word Sense Disambiguation*, WSD). Да је питање лексичке вишезначности релевантно и у савременим истраживањима српског језика показује и развој отворених ресурса који бележе значења полисемичних речи, укључујући именице, глаголе и придеве (Mišić et al. 2024).

Решавање проблема вишезначности је незаобилазно у бројним задацима обраде природног језика. У машинском превођењу различита значења исте речи могу довести до избора погрешног преводног еквивалента, што директно утиче на квалитет превода (Jurafsky and Martin 2026). У системима за претрагу информација неодређеност значења може резултирати враћањем нерелевантних докумената, јер систем није у стању да разликује различита значења истог лексичког облика у упиту или документу (Manning et al. 2008). Слични изазови јављају се и у задацима као што су анализа сентимента, где различита значења речи могу носити различиту оријентацију сентимента, као и у задацима препознавања и повезивања именованих ентитета, где је неопходно утврдити на који конкретан ентитет се одређена лексичка форма односи у датом контексту (Navigli 2009).

Због тога се у области обраде природног језика вишезначност не посматра као маргиналан проблем, већ као централни изазов семантичке анализе текста. Способност система да одреди које је значење речи употребљено у конкретном контексту представља предуслов за поуздано разумевање текста и успешну примену у различитим задацима ОПЈ-а (Navigli 2009; Jurafsky and Martin 2026). У рачунаској лингвистици овај проблем се формализује као задатак разрешења вишезначности речи, чији је циљ да се за сваку појаву вишезначне речи у тексту одреди одговарајуће значење на основу контекста.

2.3 Лексичко-семантички ресурси и Принстонски ворднет

Лексичко-семантички ресурси представљају структуриране језичке изворе у којима су лексичке јединице, њихова значења и односи међу значењима описани на експлицитан и формализован начин. Њихова улога у лингвистици и обради природног језика јесте да обезбеде организовану репрезентацију лексичког знања, односно знања о томе шта речи значе, у каквим су односима значења и како се та значења могу систематски повезивати. За разлику од приступа који се ослањају искључиво на површинске особине текста, овакви ресурси омогућавају рад на концептуалном нивоу, што је посебно важно у задацима који подразумевају тумачење контекста, разликовање значења и повезивање језичких израза са структурираним знањем (Miller 1995; "Princeton WordNet," n.d.).

Посебно значајну групу ових ресурса чине семантичке мреже, у којима су значења организована као чворови повезани формално дефинисаним релацијама. На тај начин значење није представљено изоловано, већ као део шире структуре концепата (појмова) и њихових

међусобних релација (односа). Семантичке мреже су зато важне и као практични ресурси за обраду природног језика и као концептуални модел експлицитне репрезентације значења. Управо у том оквиру развијени су ресурси типа ворднет, који су постали један од најутицајнијих модела организације лексичко-семантичког знања у савременој рачунарској лингвистици (Miller 1995; Fellbaum 1998).

Принстонски ворднет представља најпознатији и историјски најутицајнији пример таквог ресурса. Званична страница пројекта описује га као велику лексичку базу енглеског језика у којој су именице, глаголи, придеви и прилози груписани у синсете (енгл. *synset*, од *synonym set*), а синсети повезани семантичким и лексичким релацијама. Ресурс је слободно доступан за претрагу и преузимање, а изворна референца се и даље званично препоручује за цитирање, што потврђује његов трајан статус стандардног референтног извора у овој области (“Princeton WordNet,” n.d.; Miller 1995; Fellbaum 1998).

По својој организацији, Принстонски ворднет је лексичко-семантичка мрежа заснована на груписању лексема у синсете и на повезивању тих синсета релацијама. Према званичној документацији, информација у ворднету организована је око синсета, при чему сваки синсет садржи скуп литерала (енгл. *literal*), односно синонимних речи или вишечланих израза који су у неком контексту међусобно заменљиви, потом заједничку дефиницију, а често и пример употребе. Једна лексема може се појавити у више синсета и на тај начин бити повезана са више значења (“Princeton WordNet,” n.d.).

У том оквиру, синсет представља скуп литерала који деле исто значење и чине један концепт у оквиру мреже. Литерал је лексичка форма која припада конкретном синсету, односно појава леме у оквиру одређеног значења, док лема означава основни, канонски облик лексеме под којим су обједињени њени флективни облици. Појединачно лексичко значење формално се реализује кроз припадност одговарајућем синсету. Синсети су унутар ворднета међусобно повезани различитим семантичким релацијама, као што су хиперонимија/хипонимија, меронимија/холонимија и антонимија, док документација експлицитно разликује и семантичке и лексичке релације (“Princeton WordNet,” n.d.).

Оваква структура омогућава да се значење речи не посматра изоловано, већ као део уређене мреже односа међу концептима. Управо је то једна од главних предности приступа који је примењен у изградњи ворднета: експлицитно организовано лексичко-семантичко знање постаје доступно и за лингвистичку анализу и за рачунарску обраду језика. Значај Принстонског ворднета није само историјски. Новији радови показују да се он и даље користи као основа за рад са глосама⁸, за означавање значења и даље структурне анализе ресурса, што потврђује да је реч о зрелом и трајно релевантном ресурсу (Rademaker et al. 2023).

У циљу јаснијег представљања основних појмова који се користе у опису структуре ворднета, Табела 2.2 представља њихове сажете дефиниције.

Табела 2.2 Основни појмови за опис структуре ворднета

Појам	Енглески термин	Дефиниција	Ниво
Лема	<i>lemma</i>	Основни, речнички облик лексеме који обједињује њене флективне облике и може бити повезан са више синсета.	Лексички

⁸ У ворднету, глоса је кратак текстуални опис значења синсета, који може садржати објашњење, пример употребе и додатне напомене. Термин глоса користи се јер такав опис не мора имати облик потпуне лексикографске дефиниције, већ првенствено служи идентификовању и разграничењу значења у семантичкој мрежи.

Литерал	<i>literal</i>	Конкретна појава леме у оквиру одређеног синсета; лексичка форма повезана са једним значењем.	Лексички
Синсет	<i>synset</i>	Скуп литерала који деле исто значење и представљају један концепт у семантичкој мрежи.	Семантички
Значење	<i>sense</i>	Појединачно лексичко значење које се у ворднету формално реализује кроз припадност синсету.	Семантички

Лексичко-семантичка организација Принстонског ворднета, заснована на синсетима и експлицитним семантичким односима, представља полазни модел на основу кога су касније развијани ворднети за друге језике и вишејезичке семантичке мреже.

2.4 Ворднети других језика и вишејезичке семантичке мреже

После успеха Принстонског ворднета, његова парадигма је прихваћена и за друге језике, што је довело до развоја великог броја националних и вишејезичких семантичких мрежа. Један од првих и најутицајнијих таквих пројеката био је EuroWordNet, замишљен као вишејезичка база у којој су језички специфични ворднети међусобно повезани преко интерлингвалног индекса (енгл. *Inter-Lingual Index*, ILI). Овај пројекат је обухватао осам европских језика: енглески, холандски, италијански, шпански, француски, немачки, чешки и естонски. Основна идеја није била да се конструише један јединствени „универзални“ лексикон, већ да сваки језик развије сопствену мрежу синсета, док се међујезичка повезаност постиже преко заједничког концептуалног нивоа (Vossen 2004). Та архитектура је била важна јер је од почетка полазила од претпоставке да је лексичка семантика делимично заједничка, али да је њена реализација у појединачним језицима нужно језички и културно условљена.

На тој основи настао је и BalkaNet, који је применио EuroWordNet модел на балканске језике. BalkaNet је обухватао ворднете за бугарски, грчки, румунски, српски и турски, уз даље унапређивање чешког ворднета, чији је развој почео у оквиру EuroWordNet-a (Tufis et al. 2004; Krstev et al. 2004). За разлику од поједностављеног схватања да је довољно само пресликати Принстонски ворднет у други језик, искуство BalkaNet-a показало је да такво пресликавање није ни најједноставније ни најпримереније решење. Истраживачи који су радили на развоју бугарског и српског ворднета изричито наглашавају да строго један-према-један мапирање према Принстонском ворднету за ова два језика, као ни за друге језике, често није могуће, јер поједини језици лексикализују односе и појмове који у енглеском нису организовани на исти начин (Koeva et al. 2008).

Управо ту постаје јасно да међујезичко поравнање не подразумева потпуну семантичку еквиваленцију. У пракси се јављају најмање два типа проблема. Први су случајеви у којима је концепт присутан у ILI-ју, али није лексикализован у свим језицима; тада се у појединим националним ворднетима уводе такозвани нелексикализовани синсети (енгл. *non-lexicalized synsets*) ради очувања хијерархије. Други су случајеви у којима одређени појмови немају непосредан еквивалент у Принстонском ворднету, али су релевантни за више језика истог културног простора. За такве случајеве у BalkaNet-у је ILI проширен увођењем BILI (енгл. *BalkaNet Inter-Lingual Index*) концепата. У литератури се наводи да су почетни заједнички балкански концепти углавном припадали доменима као што су породични односи, религијски предмети и праксе, традиционална храна, одећа, занимања, уметност, важни догађаји и мере. Као илустративан пример наводе се две културно специфичне леме заједничке за више балканских језика: кадаиф, као назив за оријентални слаткиш, и нарицање, као назив за

обредну праксу. Ове јединице нису постојале у PWN-у, па су повезиване преко BIL I индекса (Krstev et al. 2006; Koeva et al. 2008).

За српски језик ова искуства су посебно значајна, јер је Српски ворднет настао управо у таквом вишејезичком оквиру. Према каснијем прегледу развоја ресурса, на крају пројекта BalkaNet, 2004. године, Српски ворднет је имао око 7.000 синсета повезаних са PWN-ом 2.0, а у наредним годинама додато је још приближно 14.000 синсета кроз даљи ручни и полуаутоматски рад. У једном прегледу развоја бугарског и српског ворднета наводи се и да је у раној фази дефинисано 316 српских језички специфичних концепата, насупрот 336 бугарских, 309 грчких, 545 румунских, 332 турска и 226 чешких концепата, од којих је део касније прешао у шири слој балкански специфичних појмова (Stanković, Mladenović, et al. 2018; Koeva et al. 2008). Ови подаци показују двоје: прво, да је модел ворднета довољно продуктиван да послужи као основа за дугорочан развој националних ресурса; и друго, да његова примена у вишејезичком окружењу захтева континуирано проширивање, ревизију и аутоматизацију процеса изградње и одржавања, што је посебно значајно за језике који ионако имају ограничене ресурсе.

У ширем контексту, каснији развој је додатно потврдио и прихваћеност и сложеност ворднетовог приступа. Бонд и Фостер (2013) извештавају да је њихова рана верзија *Open Multilingual* ворднета већ окупљала велике ворднете за више од 26 језика, мање скупове података за 57 језика, и више од 2 милиона значења за 117.659 концепата. Према новијој документацији OMW-а, платформа је 2023. године окупљала 33 отворена, ручно уређена ворднета, док званична страница *Global WordNet Association*⁹ данас упућује на много ширу мрежу међусобно повезаних ресурса типа ворднет, укључујући OMW као инфраструктуру која повезује ресурсе за 200+ језика. То потврђује да концепт ворднета није маргиналан или локални модел, већ један од најшире прихваћених формата за лексичко-семантичко моделовање. Међутим, истовремено остаје чињеница да број повезаних ресурса не укида проблеме семантичке нееквиваленције, различите грануларности значења и културно специфичних концепата (Global WordNet Association, n.d.; Open Multilingual Wordnet, n.d.; Bond and Foster 2013; Bond et al. 2023).

2.5 SentiWordNet и лексичко-семантичка репрезентација сентимента

Информације које се могу издвајати из текстуалних извора у најопштијем смислу могу се поделити на чињеничне исказе и исказе који изражавају ставове, оцене и емоције аутора. Чињенични искази у овом раду не представљају предмет даљег разматрања. Фокус је усмерен на исказе који садрже субјективну компоненту, односно мишљења у ширем смислу те речи.

У оквиру ове дисертације појам *мишљења* користи се као надређена категорија која обухвата ставове, вредносне процене и емоционално обојене исказе. Посебно се разматра оријентација исказа према сентименту, односно вредносна усмереност исказа као позитивна, негативна или неутрална (B. Liu 2010). У том смислу, сентимент се не изједначава са појединачним емоцијама, већ се односи на општу поларитетну вредност исказа.

Поступак аутоматске класификације таквих исказа уобичајено се означава као анализа сентимента (енгл. *sentiment analysis*) (B. Liu 2010). У аутоматској анализи сентимента уобичајено се издвајају два општа приступа. Први приступ заснива се на употреби аотираних

⁹ <https://globalwordnet.github.io/>

скупова података, у којима су текстови ручно или полуаутоматски обележени одговарајућим поларитетима, чиме се проблем анализе сентимента редукује на задатак класификације текста.

Други приступ подразумева употребу лексикона сентимента, односно ресурса у којима су лексичким јединицама придружене информације о поларитету сентимента, а који се типично развијају на основу корпуса или се изводе из других лексичких ресурса (B. Liu 2010). Израчунавање поларитета сентимента на основу значења речи, а не искључиво на основу њиховог површинског облика, може довести до веће тачности, будући да једна иста реч може имати више различитих значења са различитим оријентацијама сентимента. Полазећи од ове претпоставке, (Esuli and Sebastiani 2006) предложили су SentiWordNet (SWN) (Baccianella et al. 2010) као проширење PWN-a, у оквиру кога се поларитет сентимента дефинише на нивоу синсета, односно појединачних значења речи.

SWN се широко користи за анализу сентимента у различитим врстама текстова и доменама, укључујући блогове, рецензије производа, кратке текстуалне форме и садржаје са друштвених мрежа, како у научним истраживањима тако и у практичним применама. Такође је примењиван у анализи сентимента текстова који нису написани на енглеском (Denecke 2008), као и као полазиште за израду других лексикона сентимента. Поред тога, коришћен је и у сложенијим истраживачким контекстима, као што су откривање сарказма, интеракција човека и рачунарског агента, извлачење симболичког значења из гестова и откривање скривених знања у тексту. Судећи по броју цитата у научној литератури, SWN се често наводи као један од најчешће коришћених јавно доступних лексикона сентимента (Denecke and Reichenpfader 2023; Wankhade et al. 2022). Значајан број касније развијених лексикона следи исту структурну поставку као SWN, при чему се они могу посматрати као проширења PWN-a на конкретне језике.

У оквиру SWN-a, сваком синсету из PWN-a придружен је скуп нумеричких вредности које представљају оцене сентимента (Esuli and Sebastiani 2006). Ове оцене су формално означене као *OBJScore*, *POSScore* и *NEGScore* и односе се, редом, на степен објективности, степен позитивне оријентације сентимента и степен негативне оријентације сентимента датог значења (Baccianella et al. 2010). Вредности су реални бројеви у интервалу од 0 до 1, при чему је њихов збир за сваки синсет једнак 1. Оваква репрезентација омогућава процену да ли је дато значење објективно или субјективно, а у случају субјективности и утврђивање позитивног или негативног поларитета, при чему се сваком синсету придружује вероватноћа да у тексту изражава одговарајући сентимент.

У рачунарској имплементацији, SWN је дистрибуиран у табеларном формату, у коме сваки ред одговара једном синсету из PWN-a. Запис обухвата идентификатор синсета (врсту речи и позицију), вредности *POSScore* и *NEGScore*, листу литерала које чине синсет, као и његову дефиницију (гласу). *OBJScore* није експлицитно наведена, већ се израчунава као $1 - (POSScore + NEGScore)$, у складу са формалним ограничењем да је збир све три вредности једнак 1.

EuroWordNet и његови наследници конципирани су тако да се свако значење повезује са најближим концептом посредством ILI, који служи као посредни механизам за поравнавање значења између језика (Vossen 2004), при чему је један од таквих ресурса и Српски ворднет, о коме ће више бити речи у наставку текста. Док је у оригиналном SWN-у сентимент концептуализован као екстерно проширење постојеће структуре, у појединим националним имплементацијама, укључујући и Српски ворднет, ове вредности се моделују као унутрашња својства синсета.

Методологија SWN-а имплицитно полази од претпоставке да је оријентација сентимента инхерентна особина концепта и да се, уколико постоји поравнање синсета, она може стабилно пресликавати између језика. У пракси се, међутим, јављају две врсте ограничења. Прво, постоје структурна ограничења националних ворднет мрежа. Потпуна покривеност синсета у односу на PWN није увек обезбеђена, нити постоји потпуно једнозначно поравнање, што отежава директан пренос вредности сентимента. Други, концептуално сложенији проблем потиче од чињенице да одређени појмови немају јасан лексички или семантички еквивалент у другом језику. На пример, синсет који у српском обухвата појам „нарицање“ (**BILI-00000941**), дефинисан као „гласно изражавати жалост, запевати, тужити“, носи културолошки и прагматички специфичне конотације које немају потпуно одговарајућу репрезентацију у енглеском ворднету. У таквим случајевима, проблем није само у поравнању ресурса, већ у самој непотпуној концептуалној преклопљености између језика. Ова разлика указује да пренос вредности сентимента није искључиво техничко питање мапирања синсета, већ и питање језичке и културне специфичности концептуализације.

2.6 Методе изградње SentiWordNet-а

Изградња лексикона сентимента може се класификовати према начину доделе ознака на три општа приступа: (1) ручно означавање од стране стручњака, (2) потпуно аутоматска додела заснована на алгоритамским процедурама и моделима машинског учења и (3) полуаутоматски или хибридни приступи који комбинују ограничено ручно означавање са машинским учењем. Избор приступа условљен је факторима као што су скалабилност, трошкови израде и могућност репродукције резултата. Код ресурса великих размера, потпуно ручна израда постаје практично неодржива, што доводи до доминације аутоматских или полунадгледаних метода. У том контексту, изградња SWN-а може се посматрати као инжењерски компромис између лексикографске прецизности и алгоритамске скалабилности.

Оригинални SWN развијен је као проширење PWN-а. Оцене сентимента придружене синсетима израчунате су применом полунадгледаних метода машинског учења за класификацију текста, при чему је само мањи део скупа за учење био ручно означен, док је преостали део генерисан алгоритмима који ће бити описани у наставку текста. Добијене оцене су потом кориговане применом анализе графа, односно алгоритмом заснованом на насумичном лутању (енгл. *random walk*).

Формиране су три иницијалне групе синсета за обуку: позитивна, негативна и објективна. Полазни позитивни и негативни скупови, означени као L_p и L_n , формиран су ручним избором синсета који међу својим литералима садрже 14 „парадигматских“ позитивних и негативних термина (по седам у свакој групи), при чему су задржана само она значења која недвосмислено носе одговарајућу оријентацију поларитета сентимента (Esuli and Sebastiani 2006). Ови скупови представљају нулти корак експанзије ($K = 0$).

Експанзија позитивног и негативног скупа врши се итеративно, у K корака, при чему се у сваком кораку k генеришу скупови Tr_p^k и Tr_n^k , тако да важи:

$$Tr_p^k \supseteq Tr_p^{k-1}, Tr_n^k \supseteq Tr_n^{k-1}$$

са почетним условом:

$$Tr_p^0 = L_p, Tr_n^0 = L_n$$

Експанзија у кораку *кобухвата* два типа операција:

1. **Додавање синсета повезаних релацијама које чувају поларитет**, тј. сви синсети повезани са елементима из скупова Tr_p^{k-1} (односно Tr_n^{k-1}) релацијама као што су *similarity*, *derived from*, *pertains to*, *attribute* и *also-see*, додају се у одговарајући скуп, уколико се у њему већ не налазе.
2. **Додавање синсета повезаних релацијама које инвертују поларитет**, пре свега директном антонимијом. Синсети повезани са елементима из Tr_n^{k-1} путем антонимије додају се у Tr_p^k , и обрнуто, под условом да већ не припадају супротном скупу.

Формално, експанзија позитивног скупа може се описати као:

$$Tr_p^k = Tr_p^{k-1} \cup R_{pres}(Tr_p^{k-1}) \cup R_{inv}(Tr_n^{k-1})$$

где R_{pres} означава релације које очувавају поларитет, а R_{inv} релације које га инвертују. Аналогно важи и за негативни скуп. У сваком кораку проверава се пресек позитивног и негативног скупа. Синсети који се појаве у оба скупа третирају се као неодређени и уклањају се из Tr_p^k и Tr_n^k , након чега се могу разматрати у оквиру објективне категорије.

Објективни скуп L_o дефинисан је на другачији начин, будући да је категорија објективности комплементарна позитивној и негативној. Хеуристички је одређен као скуп синсета који:

- (a) не припадају ни Tr_p^K ни Tr_n^K ,
- (b) не садрже термине означене као позитивне или негативне у лексикону *General Inquirer* (Stone et al. 1966), лексичком ресурсу за рачунарску потпомогнуту анализу садржаја енглеских текстова, у коме су речи и поједина њихова значења сврстани у семантичке и психосоцијалне категорије, укључујући категорије позитивног и негативног вредновања.

За сваку вредност K , важи $Tr_o^K = L_o$.

Повећањем броја корака експанзије K , позитивни и негативни скупови постају све већи. Међутим, како се удаљавамо од иницијалних ручно изабраних синсета, расте вероватноћа укључивања синсета чија оријентација сентимента није недвосмислена. Другим речима, повећање K доводи до:

- повећања одзива,
- али и повећања шума у скупу за обуку, што може утицати на прецизност.

У оригиналном раду (Esuli and Sebastiani 2006) експериментално су разматране вредности $K = 0, 2, 4, 6$, при чему је показано да различите конфигурације¹⁰ утичу на баланс између прецизности и одзива, без значајних осцилација у укупној тачности (Esuli and Sebastiani 2006). Сваком синсету у скупу за обуку придружена је векторска репрезентација добијена издвајањем речи из његове глосе, уклањањем стоп-речи и мапирањем преосталих речи у нумерички

¹⁰ различите конфигурације се односе на различите итеративне скупове за тренирање и на различите моделе за машинско учење

простор. Над тако добијеним репрезентацијама примењена је TF-IDF трансформација уз косинусну нормализацију (Esuli and Sebastiani 2006). Полазна претпоставка била је да синсети сличне оријентације сентимента имају и семантички сличне глосе.

На основу ових векторских репрезентација конструисана су два бинарна класификатора: један за детекцију позитивне класе, а други за детекцију негативне класе. Први класификатор разликовао је синсети који припадају позитивној категорији од оних који јој не припадају, док је други разликовао синсети негативне категорије од осталих. Комбиновање резултата извршено је тако што су синсети класификовани као позитивни уколико су означени као позитивни и истовремено као неприпадајући негативној класи; аналогно томе, синсети су означени као негативни ако су класификовани као негативни и неприпадајући позитивној класи. Преостали синсети, односно они који нису означени да припадају иједној од ове две категорије или су означени да истовремено припадају обема, сврстани су у објективну класу.

Будући да су различите конфигурације скупова за обучавање и алгоритама машинског учења давале резултате различите, али упоредиве тачности, коначни модел организован је као ансамбл класификатора заснован на просечним оценама добијеним применом различитих комбинација скупова и алгоритама. Коришћене су четири варијанте скупова за обучавање, добијене избором различитих вредности параметра (0, 2, 4 и 6 корака експанзије), у комбинацији са алгоритмима потпорних вектора и наивног Бајеса (Esuli and Sebastiani 2006). На тај начин формирано је осам тернарних класификатора, чији су резултати агрегирани у јединствени модел и примењени на све преостале синсети у PWN-у. Добијене оцене сентимента кретале су се у интервалу од 0 до 1, са кораком од 0,125 (једна осмина). Напоменимо да децималне вредности нису поузданост оцене, већ представљају степен припадности синсета одређеној категорији сентимента, односно јачину позитивности, негативности и објективности. Након почетне класификације, примењена је додатна глобална корекција оцена заснована на структури графа. У тој фази искоришћена је чињеница да су речи у глосама PWN-а ручно означене идентификаторима синсета. Усвојена је претпоставка да постоји веза између два синсета ако глоса првог садржи реч која припада другом синсету. На тај начин PWN је представљен као граф, након чега су спроведена два итеративна процеса насумичне шетње – један за позитивне, а други за негативне оцене – при чему су вредности сентимента могле да се мењају у свакој итерацији до достизања конвергенције. Објективне вредности израчунаване су на основу позитивних и негативних оцена. Евалуација је показала да је увођење поступка насумичне шетње значајно унапредило квалитет резултата (Baccianella et al. 2010).

Резултати су евалуирани применом две методе. Поређење са лексиконом *General Inquirer* омогућило је само делимичну евалуацију, будући да се оцене сентимента у том лексикону односе на речи, а не на појединачна значења. У другој евалуацији формиран је ручно означен подскуп синсета из PWN-а, који је послужио као златни стандард под називом *Micro-WNO*, а затим су његове оцене упоређене са аутоматски добијеним оценама сентимента (Esuli and Sebastiani 2006; Formentelli et al. 2007).

Micro-WN(Op)-3.0, као адаптирана и јавно доступна верзија иницијалног *Micro-WNO* скупа, коришћена је за додатну евалуацију ресурса. Поређење синсета из SWN-а са синсетима из *Micro-WN(Op)-3.0* извршено је у односу на рангирање њихових позитивних и негативних оцена. Пошто је велики број синсета имао идентичне вредности, примењена је р-нормализована Кендалова τ дистанца (Baccianella et al. 2010). Овај приступ представљао је инспирацију за методе примењене у овом истраживању, које су детаљније описане у поглављу о методологији.

Наведене евалуације усмерене су пре свега на проверу квалитета оцена сентимента на нивоу синсета. Међутим, у литератури се лексикони сентимента засновани на SentiWordNet-у вреднују и екстерно, кроз њихову примену у задацима класификације сентимента или субјективности текста (Mohanty et al. 2017; A. Das and Bandyopadhyay 2010; Denecke 2008).

У изградњи националних лексикона сентимента заснованих на структури ворднета, уобичајена почетна стратегија јесте пресликавање постојећих оцена сентимента из SWN-а на локалну мрежу ворднета. Уколико је национални ворднет поравнат са PWN-ом посредством ILI или сличног механизма, вредности сентимента могу се директно пренети са одговарајућих синсета. У случајевима стабилног поравнања и типолошке блискости са енглеским језиком, овакво решење може дати функционално задовољавајуће резултате без значајнијих додатних интервенција. Такво пресликавање није изузетак, већ представља стандардну праксу у вишејезичким пројектима развоја ворднета. И EuroWordNet (Vossen 2004) и BalkaNet (Tufis et al. 2004) садрже пресликане вредности. У оквиру Open Multilingual WordNet-a (Bond and Paik 2012), велики број националних ворднета повезан је преко ILI механизма, што технички омогућава релативно једноставно пресликавање SWN оцена на било који поравнати ресурс. Међутим, у пракси се показало да директно пресликавање често није довољно за постизање високог степена поузданости, услед семантичких, језичких и културолошких разлика. Стога се иницијално пресликавање обично допуњује додатним поступцима проширења и валидације.

Тако су, на пример, за хинди, бенгалски и телугу језик лексикони сентимента конструисани пресликавањем оцена из SWN 3.0 и листе *Subjectivity Word List* (Wilson et al. 2005) на одговарајуће националне ворднете. Након тога, скупови су проширени употребом антонимијских релација и евалуирани кроз класификацију субјективности на корпусу и мерењем тачности поларитета на скупу од 100 ручно означених речи (A. Das and Bandyopadhyay 2010).

За синхалешки језик (енгл. *Sinhala*, ISO 639-1: *si*) SWN је комбинован са енглеско-синхалешким речником ради доделе оцена сентимента синхалешким речима и њиховим синонимима. Добијени лексикон евалуиран је на корпусу новинских текстова ручно категорисаних према оријентацији сентимента (Medagoda et al. 2015). Слично томе, *SentiTurkNet* за турски језик развијен је применом полуаутоматског поступка који комбинује пресликавање из SWN-а и ручну валидацију, уз евалуацију на ручно означеном скупу синсета (Dehkharghani et al. 2016). Слични приступи забележени су и за арапски (Alhazmi and Black 2013), вијетнамски (Vu and Park 2014), одија (Mohanty et al. 2017) и индонежански језик (Wijayanti and Arisal 2021). Заједничка карактеристика наведених приступа јесте ослањање на иницијално пресликавање из SWN-а као полазну основу, уз накнадне корекције ограниченог обима.

У односу на наведене приступе, метод предложен у овој дисертацији комбинује итеративно проширење ресурса, методе машинског учења и фино подешене језичке моделе, са циљем изградње српских лексикона сентимента шире покривености и веће поузданости.

2.7 Савремени приступи машинског учења и велики језички модели

У ранијим фазама развоја обраде природног језика текст се најчешће представљао помоћу дискретних и ручно дефинисаних особина, као што су врећа речи, n-грамски обрасци или друге статистичке карактеристике. Иако су такви приступи били функционални у бројним задацима,

они нису омогућавали компактно и семантички информативно представљање лексичких јединица.

Развој савремене обраде природног језика обележен је постепеним прелазом са статичких модела векторских репрезентација речи на моделе који значење лексеме тумаче као функцију конкретног контекста. У ранијој фази доминирали су приступи као што су Word2Vec и GloVe, у којима се свакој речи придружује једна векторска репрезентација научена на основу дистрибутивних образаца у великом корпусу (Mikolov et al. 2013; Pennington et al. 2014). Иако су ови модели представљали значајан напредак у односу на дискретне и ручно конструисане репрезентације речи, њихов недостатак је то што полисемичној лексеми додељују исту репрезентацију без обзира на значење које она има у конкретном контексту. У литератури се ово ограничење често описује као свођење више значења исте речи на један векторски приказ, што је посебно проблематично у задацима у којима је потребно прецизно разликовање значења.

Ово ограничење је непосредно релевантно за задатке који стоје у средишту ове дисертације. У анализи глоса (дефиниција) синсета није довољно да се лексема представи једним, контекстно независним вектором, јер се исте речи могу појављивати у различитим дефиницијама и носити различите семантичке улоге унутар различитих концепата. Управо зато су контекстуалне репрезентације, уведене моделима као што су ELMo и BERT, значајан корак напред: оне полазе од претпоставке да репрезентација речи мора зависити од њеног конкретног језичког окружења. ELMo моделује промене у употреби и значењу речи у различитим контекстима, обухватајући њихове синтаксичке и семантичке аспекте, док BERT уводи дубоко двосмерно представљање, при чему се репрезентација речи у свим слојевима модела изграђује узимањем у обзир и левог и десног контекста. (Peters et al. 2018; Devlin et al. 2019). На тај начин значење више није фиксна особина лексеме, већ својство које репрезентује конкретну употребу, што је од посебне важности за полисемију и за рад са кратким текстовима који садрже дефиниције, што је случај глоса у синсетима.

Архитектонску основу већине савремених великих језичких модела чини трансформер, представљен у раду *Attention Is All You Need* (Vaswani et al. 2017), који у тренутку писања овог рада има више од 265.000 цитата према академском претраживачу *Google Scholar*. За разлику од рекурентних секвенцијалних архитектура, трансформер не обрађује текст строгим редом, већ улазни низ представља као скуп токена чије се међусобне везе моделирају механизмом пажње. Улазни текст се најпре сегментира на токене, након чега се сваком токenu придружује векторска репрезентација. Пошто сам механизам пажње не садржи уграђен појам редоследа, трансформер уводи позиционо кодирање како би модел могао да узме у обзир и положај сваког токена у секвенци. Тако се текст не представља само као скуп јединица, већ као уређена секвенца у којој су и идентитет и позиција елемената релевантни за даљу обраду.

Централна операција трансформера јесте механизам самопажње (енгл. *self-attention*), посебан случај механизма пажње (енгл. *attention*) којим се репрезентација сваког токена израчунава у односу на све остале токене у истој секвенци. Формално, механизам пажње може се записати на следећи начин:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

где Q означава матрицу упита¹¹ (енгл. *queries*), K матрицу кључева (енгл. *keys*), V матрицу вредности (енгл. *values*), а d_k димензију кључева (Vaswani et al. 2017). Интуитивно посматрано, ова операција омогућава моделу да за сваки токен одреди који су други токени у секвенци најрелевантнији за формирање његове репрезентације. Уместо да се контекст гради искључиво кроз локалне или секвенцијалне прелазе, као код рекурентних модела, трансформер омогућава директно повезивање удаљених елемената унутар исте дефиниције¹². Ово својство је додатно унапређено вишеглавим механизмом пажње (*multi-head attention*), у коме се пажња рачуна кроз више паралелних пројекција, тако да модел може истовремено да захвати различите типове односа између токена. Узастопним слагањем више трансформерских блокова добијају се све апстрактније репрезентације, које постепено прелазе од површинских образаца ка сложенијим синтаксичким и семантичким зависностима.

Значај механизма пажње за семантичку анализу нарочито долази до изражаја у кратким, али семантички густим текстовима. Глосе синсета по правилу су кратке, али у сажетом облику садрже више врста информација: упућују на надређени појам, издвајају својства по којима се појам разликује од сродних појмова, одређују област или услове употребе, а понекад садрже и прагматичке напомене. Тако је у Српском ворднету „сусед“ одређен као „особа“, а „разбој“ као „гимнастичка справа“, чиме се наводе њихови надређени појмови. Ове одреднице додатно се прецизирају навођењем својстава која их разликују од осталих: сусед је особа која „живи (или је смештена) близу друге“ особе, док се разбој описује као справа која се „састоји од две дрвене летве које придржавају носачи“. Поједине глосе садрже и ограничења употребе, попут напомена „о биљкама“ уз придев „зимзелен“, „у физици“ уз „светлост“ и „у рачунарству“ уз „погон“. Прагматичке информације налазе се, на пример, у глоси синсета „руља, светина, гомила“, дефинисаног као „погрдни изрази за обичан народ“, где придев „погрдни“ означава вредносни однос који се исказује употребом тих назива. Управо зато је важно да модел може да узме у обзир све делове дефиниције истовремено, а не само локално суседство речи. За разлику од рекурентних неуронских мрежа, укључујући архитектуре типа LSTM, код којих се информација преноси кроз низ узастопних стања, трансформер омогућава непосредну интеракцију било која два положаја у секвенци. LSTM је уведен управо да би ублажио проблем дугорочних зависности у рекурентним мрежама, али трансформер ту врсту ограничења решава другачије, одустајањем од саме секвенцијалне рекуренције и ослањањем на пажњу као примарни механизам моделирања зависности (Hochreiter and Schmidhuber 1997; Vaswani et al. 2017). За глосе, које су кратке, али концептуално концентрисане, ова способност глобалног моделирања односа између свих токена представља посебну предност.

У ширем методолошком смислу, велики језички модели могу се посматрати као део парадигме базних модела (енгл. *foundation models*), односно модела обучених на широкој и хетерогеној грађи, који се потом могу прилагођавати различитим низводним задацима. Ова парадигма је повезана са учењем преносом знања (енгл. *transfer learning*) и са могућношћу да један претходно обучени модел обезбеди општу језичку основу за више различитих задатака, укључујући класификацију, генерисање текста и семантичко поравнавање (Bommasani et al. 2021). Радови посвећени моделима као што је GPT-3 додатно су показали да велики језички модели, захваљујући својој величини, обиму података за обучавање и примењеним поступцима обуке, могу да уче нове задатке на основу малог броја примера (енгл. *few-shot*

¹¹ У овом контексту термин „упит“ не означава језички упит, већ вектор који представља информацију коју текући токен тражи у контексту. Кључеви описују својства на основу којих се остали токени пореде са упитом, док вредности носе садржај који се, у складу са израчунатим тежинама пажње, укључује у нову репрезентацију токена.

¹² дефиниција је улазна секвенца, а упит вектор који се посебно образује за сваки њен токен

learning) и учења из контекста (енгл. *in-context learning*), и то без експлицитног финог подешавања за сваки појединачни задатак (T. Brown et al. 2020). У контексту ове дисертације то је важно зато што ВЈМ могу да послуже као методолошки инструмент у задацима анализе, класификације и делимично контролисаног генерисања текста, примењених на глосама и синсетима. Ипак, ова употреба не подразумева да они могу заменити експлицитно структурирано лексикографско знање.

Овде је потребно јасно нагласити методолошку границу. ВЈМ, иако изузетно моћни у моделирању језичких образаца, не поседују експлицитну концептуалну структуру какву обезбеђују лексичко-семантички ресурси типа ворднет. Лексичке мреже организују значења као дискретне, релационо повезане јединице, док су репрезентације ВЈМ пре свега дистрибутивне и имплицитне. Управо због те разлике савремена литература не посматра лексичке ресурсе и моделе ОПЈ-а као међусобно замениве, већ као комплементарне изворе знања: методе ОПЈ-а могу помоћи изградњи и проширивању лексичких ресурса, а сами лексички ресурси могу побољшати рад система за ОПЈ (Camacho-Collados and Pilehvar 2018). И у новијим радовима о ворднету у доба великих језичких модела показује се да дефиниције, примери и лексикографска структура и даље задржавају централну улогу и у условима доминације ВЈМ, управо зато што обезбеђују експлицитну семантичку организацију која недостаје чисто дистрибутивним моделима (Almeman et al. 2024).

Осим тога, генеративни модели су подложни халуцинацијама, односно производњи уверљивог, али непоткрепљеног или фактички нетачног садржаја. Ова појава је систематски описана у прегледној литератури о генеративним моделима и представља једно од кључних ограничења њихове примене у задацима који захтевају поуздану лексичко-семантичку контролу (Ji et al. 2023). Поред тога, њихово понашање у великој мери зависи од карактеристика података и задатака на којима су претходно обучени, као и од начина на који су накнадно прилагођени конкретном проблему (Bommasani et al. 2021). Зато се у овом раду ВЈМ не третирају као замена за ворднет, већ као комплементаран инструмент унутар ширег лексикографског и семантичког система. Њихова вредност није у томе што укидају потребу за експлицитним ресурсом, већ у томе што могу да допуне тај ресурс бољим моделирањем контекста, флексибилнијом обрадом кратких дефиницијских текстова и ефикаснијим повезивањем структурисаног знања са реалном језичком употребом. Та позиција је у потпуности у складу са основном аргументацијом ове дисертације, у којој се ВЈМ посматрају као снажан, али методолошки ограничен алат чија пуна употребна вредност долази до изражаја тек у споју са експлицитно моделираним лексичко-семантичким знањем.

Због тога је приликом примене великих језичких модела у овој дисертацији важно разликовати фазе њиховог прилагођавања и употребе, пре свега фино подешавање, фазу закључивања (инференцу) и поступке којима се понашање модела усмерава ка конкретном задатку.

2.8 Фино подешавање и закључивање

Савремени језички модели развијају се кроз више узастопних фаза које обухватају претходну обуку (енгл. *pretraining*), фино подешавање (енгл. *fine-tuning*) и закључивање (енгл. *inference*) (Jurafsky and Martin 2026; Devlin et al. 2019; Grattafiori et al. 2024). У фази претходне обуке модели се обучавају на веома великим корпусима текста са циљем усвајања општих језичких карактеристика и репрезентовања статистичких правилности (закономерности) језика. Резултат овог процеса су претходно обучени модели који се могу примењивати у широком спектру задатака обраде природног језика.

Иако се сама претходна обука не спроводи у оквиру ове дисертације, она је значајна за избор модела, будући да различити модели могу бити обучени на различитим типовима и количинама података. У пракси се стога врши избор између општих вишејезичких модела, модела обучених за појединачне језике, као и модела који су додатно обучени на сродним језицима или специфичним доменима. Тај избор може имати значајан утицај на применљивост модела у задацима обраде српског језика.

Када постоје довољно велики и поуздано означени скупови података, претходно обучени модели могу се додатно прилагодити конкретном задатку кроз фазу финог подешавања. Овај поступак омогућава да се опште језичко знање модела прилагоди специфичним задацима. У овој дисертацији фино подешавање се примењује у задацима у којима су доступни означени подаци, као што је класификација сентимента синсета у оквиру ресурса Српски ворднет.

Поред тога, претходно обучени модели могу се користити и директно у фази закључивања (инференције), односно примене већ обученог модела на нове податке без додатне обуке. У овом контексту закључивање може обухватати различите типове задатака, као што су класификација, избор одговарајућег значења или генерисање текстуалног излаза. Овај приступ је посебно релевантан у задацима у којима је количина означених података ограничена, као што су разрешење вишезначности речи (енгл. *Word Sense Disambiguation*) или анализа и превођење лексикографских описа значења.

У таквим сценаријима понашање модела у великој мери зависи од начина формулисања задатка и контекста који се моделу задаје. Због тога у новијим приступима посебну улогу има обликовање промптова и њихово прослеђивање моделима, познато као инжењерство промптова (енгл. *prompt engineering*), које омогућава систематско конструисање инструкција и контекста ради усмеравања понашања великих језичких модела у фази закључивања. Следећи одељак разматра принципе инжењерства упита и његову примену у задацима обраде лексичко-семантичких ресурса.

2.9 Обликовање промптова моделима

Инжењерство промптова¹³ (енгл. *prompt engineering*) последњих година се издвојило као практично и методолошки релевантно средство за прилагођавање великих језичких модела (ВЈМ) конкретним задацима без додатног финог подешавања на самом задатку. Основна идеја овог приступа јесте да се понашање модела усмерава пажљиво обликованим исказима на природном језику који експлицитно одређују циљ, ограничења и очекивани излаз. На тај начин задаци се могу решавати без примера или уз мали број примера, односно у режиму без примера и у режиму са малим бројем примера (енгл. *zero-shot* и *few-shot*) (Т. Brown et al. 2020).

Такав приступ постао је нарочито значајан са развојем великих генеративних модела, будући да се показало да исти модел може извршавати различите задатке на основу инструкције и контекстуалних примера, без посебне додатне обуке за сваки појединачни задатак (Т. Brown et al. 2020). У најширем смислу, инжењерство промптова обухвата избор и распоред контекстуалних информација, формулацију правила, задавање улоге моделу, увођење примера, као и спецификацију излазног формата. Рани радови показали су да модели могу бити веома

¹³ У недостатку устаљеног термина у српском језику, разматрани су изрази инжењерство/обликовање упита, обликовање инструкција и обликовање налога. Ипак, ради прецизнијег означавања појма у контексту генеративне вештачке интелигенције и избегавања двосмислености коју могу изазвати речи упит, инструкција и налог, задржан је англицизам промпт.

осетљиви на измене у формулацији инструкција и начину представљања задатка, због чега упит не треба посматрати као неформалну корисничку поруку, већ као експериментално контролисан улаз у систем (Reynolds and McDonell 2021). У том смислу, инжењерство промптова представља поступак систематског обликовања интерфејса између истраживачког задатка и генеративног модела. У новијим прегледним радовима ова област је додатно систематизована кроз типологије техника и препоруке за њихову примену (Schulhoff et al. 2024).

Истраживања указују на то да је ефикасност промпта у великој мери условљена начином уоквиравања задатка. Експлицитно задавање улоге моделу може повећати кохерентност и доменску усмереност одговора, док обрасци који унапред прописују излазни формат, на пример фиксна поља, табеларну структуру или строго дефинисан синтаксни образац, доприносе смањењу разноликости и спонтане генерације која није погодна за даљу машинску обраду (Reynolds and McDonell 2021; Schulhoff et al. 2024). Због тога је оправдано да промпт садржи јасно разграничене компоненте: (а) спецификацију задатка, (б) правила за избор одговора, (в) ограничења која искључују нежељене интерпретације и (г) експлицитну инструкцију о формату повратног излаза, како би се резултати могли систематски рашчланити, вредновати и упоређивати (Schulhoff et al. 2024).

Посебно важан облик овог приступа представља уланчавање промптова (енгл. *prompt chaining*), то јест разлагање сложеног задатка на више међусобно повезаних корака, при чему излаз једног корака постаје улаз, контекст или ограничење у наредном кораку. Такав поступак је користан онда када један јединствени промпт не обезбеђује довољну контролу над процесом, или када сама природа задатка захтева фазно извођење. Иако се овај термин у литератури не употребљава увек потпуно уједначено, радови из области инжењерства промптова и вишестепеног усмеравања модела показују да разлагање задатка на низ повезаних корака може допринети већој стабилности и употребљивости излаза, посебно када је потребно раздвојити анализу, проверу и коначно форматирање одговора (Madaan et al. 2023; Reynolds and McDonell 2021; Schulhoff et al. 2024). У контексту ове дисертације то је нарочито релевантно у случајевима у којима се сложена структура синсета, дефиниција, примера и лексичких односа не може поуздано обрадити у једном кораку, већ захтева поступно вођење модела кроз сродне подзадатке. Конкретни облици таквог поступка биће описани у методолошком делу рада.

Када је реч о задацима који захтевају сложеније закључивање, значајан корпус радова показује да подстицање модела на образложено одговарање може повећати поузданост одлуке и омогућити накнадну анализу грешака (Wei et al. 2022). Ипак, у многим применама није неопходно захтевати потпуно експлицитно и опширно образлагање кроз више корака; често је довољна умерена форма образложења, односно кратко навођење разлога за избор, које повећава разумљивост без непотребног усложњавања излаза. У том смислу корисно је разликовати поступак закључивања као интерни процес модела од образложења као експлицитно исказаног излаза, што је важно и у ширем оквиру објашњиве обраде природног језика (Wiegrefe and Marasovic 2021). Сродно томе, приступи у којима систем реагује на свој одзив и дорађује се (енгл. *self-feedback* and *self-refinement*), у којима модел накнадно процењује и коригује сопствени одговор, показали су потенцијал за стабилизацију резултата, мада њихова примена подразумева додатне итерације и зато мора бити методолошки оправдана у односу на циљеве и трошкове експеримента (Madaan et al. 2023). Иако је у ранијим радовима експлицитно подстицање модела на образлагање у више корака често доносило побољшања, код новијих модела са развијенијим унутрашњим механизмима закључивања потреба за

оваквим приступом постаје мање изражена, па се тежиште све више помера са експлицитног приказивања корака на јасно формулисање задатка, ограничења и излазног формата.

Инжењерство промптова није релевантно само за разрешење вишезначности речи, већ и за друге задатке у којима је неопходно прецизно усмерити модел и контролисати излаз. У анализи сентимента, на пример, формулација промпта утиче на то да ли ће модел задатак тумачити као грубу класификацију поларитета, финију процену интензитета, идентификацију аспекта¹⁴ или издвајање доказа за донету оцену. Новија емпиријска истраживања показују да ВЈМ могу постићи добре резултате у једноставнијим задацима анализе сентимента, али да заостају у сложенијим сценаријима који захтевају дубље разумевање или структурирану информацију о сентименту; истовремено, њихов потенцијал постаје нарочито видљив у режиму са малим бројем примера, када су анотирани ресурси ограничени (Zhang et al. 2024).

Слично томе, у машинском превођењу ВЈМ могу производити употребљиве преводе само на основу инструкције, али њихова поузданост расте када су промптови пажљиво структурирани, када садрже термилошка ограничења или када се поступак превођења разложи на више корака. То је нарочито важно код језика са мање ресурса, где ограничена заступљеност ресурса на неком језику у предобучавању¹⁵ може умањити стабилност излаза. Радови који се баве превођењем језика са ограниченим ресурсима показују да стратегије са више корака, укључујући увођење експлицитних граматичких и синтаксичких образаца у промпт, могу значајно унапредити квалитет превода у односу на једноставан приступ без примера (Guo et al. 2024).

Посебан изазов представља примена инжењерства промптова у вишејезичном окружењу и у раду с језицима с недовољно ресурса. Иако се показало да велики модели могу дати употребљиве резултате и за слабије заступљене језике, њихове перформансе често зависе од доступности квалитетних примера, стабилности термилошког пресликавања и пажљиво контролисаног односа између изворног и циљног језика унутар самог промпта (Guo et al. 2024). Због тога је пожељно избегавати непотребне трансформације унутар промпта, као што су посредно превођење или неконтролисано мешање језика, када то није неопходно за дефинисање задатка. Поред тога, код дужих улаза може се јавити неравномерна расподела пажње, при чему се делови садржаја у средини улаза занемарују, што има непосредне импликације за форматирање промптова који садрже више кандидате за дефиниције, примере или ограничења. То мотивише употребу јасних структурних сигнала, као што су нумерација, сегментација и експлицитне ознаке поља¹⁶ (N. F. Liu et al. 2024).

У односу на фино подешавање, као поступак трајнијег прилагођавања параметара модела, инжењерство промптова примењује се у фази примене модела, односно закључивања, и

¹⁴ Под аспектом се подразумева реч или речи које су тригер (сигнал) за додељивање сентимента.

¹⁵ Количина доступне текстуалне грађе на српском језику знатно је мања него на великим светским језицима. Предобучавање подразумева обучавање модела на великим количинама необележеног текста, без непосредног решавања конкретног изводног задатка. Данас је већина модела вишејезична и обучава се на текстовима на великом броју језика, па мања заступљеност српског језика у подацима за предобучавање свакако може утицати на квалитет резултата које модел генерише на српском језику.

¹⁶ Код задатака класификације важно је да промпт садржи дефиниције класа кандидата, одговарајуће примере и ограничења. На пример, у задатку препознавања именованих ентитета моделу треба објаснити шта се у конкретној схеми сматра особом, локацијом или догађајем. Корисно је додати и примере граничних случајева: „Свети Сава“ може означавати особу када се говори о светитељу, а догађај када се односи на празник или славу; слично томе, писање речи „Сунце“ великим почетним словом може указивати на назив небеског тела, док „сунце“ написано малим словом не мора бити именовани ентитет. Ако упит садржи више кандидата дефиниција, примера и ограничења, потребно их је јасно структурирати: дефиниције, примере и ограничења треба раздвојити у посебне одељке, сваки одељак означити, а кандидате нумерисати.

омогућава усмеравање понашања система без измене самог модела. У целини, инжењерство промптова омогућава добијање контролисаног, разумљивог и машински обрадивог излаза које не захтева додатну обуку на задатку, али је његова успешност тесно повезана са формалном структуром промпта, јасноћом инструкција, организацијом поступка у више корака када је то потребно и доследном контролом излазног формата (Schulhoff et al. 2024). У овој дисертацији такав приступ је посебно релевантан у задацима у којима се од модела очекује контролисан и проверљив излаз, као што су анотација значења, превод синсета и поравнање лексичко-семантичких ресурса.

2.10 Генерисање података и дестилација модела

Велики језички модели не користе се само као системи за непосредно решавање задатака, већ и као важан методолошки алат у припреми података за обуку других модела. Њихова значајна улога огледа се у генерисању синтетичких података, аутоматском означавању неанотираних примера и формирању скупова погодних за обуку мањих, задатку прилагођених модела (Y. Wang et al. 2023). У условима ограничених језичких ресурса, као и у задацима и доменама за које не постоје довољно развијени скупови података, овакав приступ је посебно важан, јер је количина ручно означеног материјала ограничена, док је непосредна примена великих модела на већим колекцијама текстова често рачунарски и организационо захтевна (Y. Wang et al. 2023).

У том контексту, велики језички модели могу послужити као средство за проширење постојећег скупа података новим примерима или за аутоматско додељивање ознака већ постојећим, неанотираним примерима. За потребе овог истраживања посебно је важан други приступ, односно употреба великог језичког модела за аутоматско или полуаутоматско означавање података, при чему модел производи тражене ознаке на основу пажљиво формулисаних промптова. Тако добијени скуп може се затим употребити за обуку мањег модела, на пример модела заснованог на архитектури BERT, који је погоднији за локалну примену, брже закључивање и стабилније експериментално испитивање (Devlin et al. 2019). Поред тога, овакав приступ омогућава и формирање већег скупа за евалуацију у задацима за које такав ресурс још не постоји или није довољно развијен, чиме се смањује ослањање искључиво на људске експерте за обављање анотације, што је по правилу временски, организационо и финансијски веома захтевно (Y. Wang et al. 2023).

Предност овог приступа огледа се у томе што омогућава релативно брзо добијање већег броја означених примера, што је нарочито важно у задацима где је ручна анотација скупа, спора или захтева висок ниво лингвистичке стручности. На тај начин велики језички модел не функционише само као крајњи извршилац задатка, већ и као извор надзорног сигнала за обуку компактнијег и ефикаснијег модела.

Међутим, овакав поступак носи и одређене ризике. Квалитет добијеног скупа зависи од формулације промпта, од изабраног модела и од јасноће циљних категорија. Поред тога, велики језички модели могу производити нетачне или непоуздане излазе, као и репродуковати пристрасности присутне у подацима на којима су обучени. Због тога се аутоматски означени подаци не могу некритички прихватити, већ захтевају одговарајућу контролу квалитета, као што су филтрирање, провера формалне исправности, поређење више генерисања и ручна провера ограниченог узорка.

Овај приступ може се довести у везу са широм парадигмом дестилације знања, односно преноса знања са већег на мањи модел. Модел са којег се знање преноси назива се наставнички модел (енгл. *teacher model*), а модел који то знање усваја студентски модел (енгл. *student model*). У класичном облику, дестилација подразумева да студентски модел учи из вероватноћа са излаза наставничког модела, односно из такозваних меких циљева¹⁷ (енгл. *soft targets*) (Hinton et al. 2015).

За ово истраживање релевантнији је посредни облик преноса знања, у коме велики језички модел најпре производи или предлаже ознаке за већи број примера, а затим се на тако добијеном скупу обучава мањи модел. Другим речима, знање великог модела преноси се посредством аутоматски означеног скупа података, а не директним усклађивањем унутрашњих репрезентација или расподела вероватноћа. У том ужем смислу, овде је ближи појам дестилација модела, али не у строго техничком облику преноса меких циљева, већ као практичан пренос корисног понашања великог модела на мањи модел путем аутоматски означеног скупа.

За истраживања усмерена на српски језик овакав поступак има посебан значај, јер омогућава да се ограничени ручно означени ресурси допуне или прошире, а да се затим обучи компактнији модел погодан за даљу примену. У том смислу, генерисање података и аутоматско означавање помоћу великих језичких модела представљају важан методолошки механизам за изградњу нових ресурса и за припрему података за обуку ефикаснијих модела у условима ограничених ресурса.

2.11 Ресурси и приступи за разрешење вишезначности речи

Задатак аутоматског разрешења вишезначности речи, који је уведен у одељку 2.2 као последица вишезначности лексема у природном језику, представља један од централних проблема рачунарске семантике. Основно питање које овај задатак поставља јесте како аутоматски одредити које од више могућих значења неке лексеме одговара конкретној употреби у датом контексту.

Током развоја ове области предложен је велики број приступа и ресурса за решавање овог проблема. Уопштено, постојеће методе могу се груписати у неколико ширих категорија, међу којима се најчешће издвајају приступи засновани на знању и приступи засновани на машинском учењу.

Приступи засновани на знању ослањају се на структуриране лексичко-семантичке ресурсе, као што су PWN (Fellbaum 1998) или BabelNet¹⁸ (Navigli 2009), и третирају проблем разрешења као задатак избора најприкладнијег значења из унапред дефинисаног инвентара синсета. Ране методе користиле су дефиниције (глосе) из машински читљивих речника као основни извор информација. Класичан пример представља Лесков алгоритам (Lesk 1986), који се заснива на

¹⁷ Генеративни модел не располаже само једним могућим одговором, већ при сваком кораку генерисања израчунава расподелу вероватноће за наредни токен. Најчешће се бира највероватнији токен, док параметар температуре одређује у којој мери ће бити разматране и мање вероватне могућности. Код тврде дестилације користи се само коначан одговор или ознака наставничког модела, док се код меке дестилације користе и његова расподела вероватноће, односно *логити* (енг. *logits*). Логити су, дакле, сирове, ненормализоване нумеричке вредности које модел израчунава за сваки могући наредни токен. Оне се помоћу функције *softmax* претварају у расподелу вероватноћа.

¹⁸ Вишејезичка семантичка мрежа и лексичко-семантички ресурс који интегрише податке из Ворднета, Википедије и других извора, повезујући речи и њихова значења у великом броју језика, <https://babelnet.org/>.

мерењу преклапања између глосе кандидатског значења и контекста у којем се реч јавља. Овај приступ полази од претпоставке да ће исправно значење делити више заједничких појмова са датим контекстом него остала значења.

Са развојем семантичких мрежа, пажња се помера са локалног преклапања дефиниција ка глобалној структури лексичког графа. Модернији приступи засновани на знању користе технике графа, попут мерења централности, повезаности чворова или алгоритама шетње кроз граф, како би моделирали односе између значења (Agirre et al. 2014; Cao et al. 2025). У овим моделима разрешење се не заснива искључиво на површинским карактеристикама текста, већ на позицији кандидатског значења у широј семантичкој структури.

Иако приступи засновани на знању омогућавају већу разумљивост процеса одлучивања и не зависе од великих означених корпуса, надгледани модели су временом показали боље перформансе у условима доступности аотираних података. У надгледаном парадигматском оквиру проблем разрешења вишезначности речи се формулише као класификациони задатак, где је циљ да се научи мапирање између контекстуалне употребе речи и одговарајуће ознаке значења. Рани системи, често означавани као модели стручњака за речи (лема-специфични модели) (енгл. *word expert models*), користили су ручно осмишљена обележја специфична за сваку полисемичну реч. Међутим, надоградња је у оваквом приступу ограничена, јер захтева одвојено моделирање за сваку лему.

Суштинска разлика између ова два парадигматска оквира огледа се у концепту значења који они имплицирају. Док приступи засновани на знању полазе од експлицитно дефинисаних и стабилних инвентара значења, надгледани модели значење третирају као категорију која се учи из података. У оба случаја, међутим, централну улогу има унапред дефинисана структура значења, најчешће заснована на инвентарима ворднета, што омогућава да се задатак формализује као избор једног од дискретних синсета.

Овај класични парадигматски оквир поставља теоријску основу за касније хибридне приступе и савремене моделе засноване на контекстуалним језичким репрезентацијама. Истовремено, он јасно указује на ограничења која постају израженија у језицима са ограниченим ресурсима, где недостају обимни ручно означени корпуси или потпуно развијени лексички ресурси. Управо у том простору настаје потреба за новим методолошким решењима, која комбинују структурално знање и дистрибутивне репрезентације, што ће бити размотрено у наставку.

Појава ВЈМ и учење преносом знања значајно су изменили парадигму разрешења вишезначности речи. Док су класични надгледани приступи захтевали обимне ручно означене корпусе за сваки појединачни језик и домен примене, савремени модели омогућавају пренос знања између језика и задатака, чиме се смањује зависност од локалних ресурса.

Показано је да контекстуалне репрезентације у претходно обученим моделима већ интерно одражавају разликовање значења, чак и без финог подешавања за специфичан задатак (Loureiro et al. 2021). У таквом приступу разрешење вишезначности речи може се формулисати као проблем мерења сличности у простору векторских репрезентација, при чему се значење одређује на основу близине контекстуалне репрезентације и репрезентација кандидатских значења. Ови резултати представљају концептуални мост између класичних надгледаних метода и приступа заснованих на ВЈМ.

Истовремено, развијени су обимни вишејезички оквири за евалуацију разрешења вишезначности речи. XL-WSD представља први обједињени вишејезички оквир великог обима, који обухвата 18 језика и више од 99 000 златно означених значења према референтном

стандарду (назива се и златни стандард, енгл. *gold standard*) (Pasini et al. 2021). Резултати показују да вишејезички модели постижу успешан пренос знања у режиму без примера када се обучавају искључиво на енглеским подацима, често надмашујући класичне приступе засноване на знању. Ипак, перформансе значајно опадају код језика са мање развијеним ресурсима, као и код ретких значења речи и код речи са много значења, што је нарочито изражено код глагола. Ово указује на ограничења учења преносом знања у условима структурне и типолошке различитости.

Поред оваквих вишејезичких евалуационих оквира, развијени су и ресурси који омогућавају прецизније повезивање лексичко-семантичких инвентара са реалним контекстуалним употребама речи. Један од таквих ресурса је ELEXIS-WSD, вишејезички паралелни, ручно анотирани корпус за разрешење вишезначности речи, осмишљен тако да повезује контекстуалне појаве пунозначних речи са отвореним инвентарима значења у више европских језика. У почетној фази развоја ресурс је био конципиран за десет европских језика: бугарски, дански, холандски, енглески, естонски, мађарски, италијански, португалски, словеначки и шпански. Његова карактеристична одлика је паралелна структура, која омогућава поређење избора значења кроз више језика у истим или еквивалентним контекстима. За разлику од традиционалних једнојезичких WSD корпуса, ELEXIS-WSD омогућава анализу интерлингвалне конзистентности значења и представља важну основу за развој и евалуацију вишејезичких и трансферних приступа разрешењу вишезначности речи (Martelli et al. 2023, 2021). У новијим фазама развоја корпус је даље прошириван и обогаћиван новим поткорпусима и слојевима анотације, што потврђује његов значај као основе за вишејезичка и трансферна истраживања разрешења вишезначности речи (Čibej et al. 2025).

Поред корпуса који повезују контекстуалне појаве речи са инвентарима значења, развијени су и ресурси који полазе од дефиниција као примарног носиоца семантичке информације. У том контексту издваја се SenseDefs, вишејезични корпус аутоматски анотираних дефиниција, који обухвата десетине милиона дефиниција у великом броју језика. Овај ресурс конструисан је комбиновањем више лексикографских и енциклопедијских извора у оквиру јединственог вишејезичног инвентара значења, заснованог на BabelNet-у. Основна идеја приступа јесте да се дефиниције истог појма у различитим језицима користе као проширени контекст за разрешење вишезначности, чиме се превазилази ограничење кратких и често недовољно информативних глоса. Додатно, у поступак анотације укључене су и дистрибутивне репрезентације, чиме се комбинују структурно знање и статистичка сличност. Резултат је велики корпус семантички повезаних дефиниција, који се може користити као основа за даља истраживања у разрешењу вишезначности, изградњи семантичких ресурса и задацима екстракције знања (Camacho-Collados et al. 2019).

Посебан правац истраживања обухвата стратегије засноване на промптовима, у којима се разрешење вишезначности формулише као класификациони задатак у формату промпт–одговор. Показано је да неструктурисани упити дају слабе резултате, док укључивање глоса, филтрирање значења кандидата по врсти речи и додавање илустративних примера значајно побољшавају тачност (Sumanathilaka et al. 2024). Најбољи резултати постижу се комбинацијом режима са малим бројем примера и експлицитног вођења модела кроз процес резоновања. Ови налази указују да, иако ВЈМ интерно кодирају информације релевантне за значење, успешно разрешење вишезначности захтева експлицитна лексичка ограничења и структурну контролу задатка.

Систематске анализе метода разрешења вишезначности речи за морфолошки богате језике и језике са мање развијеним ресурсима указују на трајне изазове, укључујући зависност од

обимних обележених корпуса и ограничену преносивост модела обучених на језицима са добро развијеним ресурсима (Masethe et al. 2024). За српски језик, где не постоји стандардизовани евалуациони оквир за разрешење вишезначности речи, ови проблеми су посебно изражени.

Показано је и да ВЈМ могу вршити избор значења у режиму без примера, без експлицитне обуке на означеним подацима (Basile et al. 2025). Међутим, резултати постигнути за више европских језика указују на неуједначену стабилност и ограничену преносивост на типолошки различите и мање заступљене језике. Ово отвара питање у којој мери се способности разрешења вишезначности могу поуздано применити на српски језик и у којим условима је неопходна интеграција са постојећим лексичким ресурсима.

Утицај ВЈМ на разрешење вишезначности речи стога се не огледа искључиво у повећању перформанси, већ у померању методолошког оквира: од строго надгледаних система ка моделима који комбинују учење преносом знања, експлицитно лексичко знање и контролисано упутно вођење. Ова промена је од посебног значаја за језике са мање развијеним ресурсима, где класични надгледани приступи често нису изводљиви у пуном обиму.

Посебно релевантан низводни контекст за разрешење вишезначности речи јесте машинско превођење, јер погрешан избор значења полисемичне речи може довести до погрешног избора преводног еквивалента. Ова веза је препозната већ у раним разматрањима машинског превођења, у којима је лексичка вишезначност издвојена као један од централних изазова аутоматске интерпретације текста (Weaver 1949). У каснијим статистичким системима испитивано је да ли се разрешење вишезначности речи може искористити као спољна информација за избор преводног еквивалента, али резултати нису били једнозначни. Резултати неких истраживања показивали су да засебан модел за разрешење вишезначности није нужно довољан за побољшање статистичког машинског превођења, док су други радови указивали да корист зависи од начина на који се информација о значењу или преводна информација интегрише у сам преводилачки систем (Chan et al. 2007; Carpuat and Wu 2005; Vickrey et al. 2005).

У неуронском машинском превођењу¹⁹ (енгл. *neural machine translation* (NMT)) овај проблем добија нешто другачији облик. Пошто модели засновани на неуронским мрежама већ интерно користе контекстуалне репрезентације, питање више није само да ли је потребно разрешити вишезначност речи у контексту, већ да ли експлицитна информација о изабраном значењу доноси додатни добитак у односу на оно што модел већ уме да изведе из контекста. У досадашњим истраживањима испитивани су различити начини увођења информације о значењу у моделе засноване на неуронским мрежама, у поступак декодирања или у обраду хомографа (Rios Gonzales et al. 2017; Pu et al. 2018; Frederick Liu et al. 2018). Касније евалуације преводилачких система на тестовима усмереним ка лексичкој вишезначности указују на побољшање савремених система, али и на методолошку осетљивост процене доприноса експлицитног разрешења вишезначности (Rios et al. 2018; Raganato et al. 2019; Scherrer et al. 2020).

Овај преглед одређује методолошко место експеримента у овој дисертацији. Ранији радови углавном испитују интеграцију информације о значењу у модел, у поступак обучавања, у декодирање или у специјализовану евалуацију двосмислених речи. Насупрот томе, овде се испитује једноставнија и практично доступнија стратегија: да ли кратка напомена о значењу, изведена из инвентара ворднета и додата улазној реченици, може усмерити већ постојећи

¹⁹ Машинском превођењу у коме се користе модели засновани на вештачким неуронским мрежама.

преводилачки модел без његовог додатног обучавања. Због тога резултате ове провере треба тумачити као испитивање ограничене стратегије употребе информације о изабраном значењу, а не као општу оцену вредности разрешења вишезначности речи за машинско превођење.

2.12 Српски ворднет: стање, значај и изазови

У оквиру ширег развоја вишејезичких семантичких мрежа, Српски ворднет представља најзначајнији доступни лексичко-семантички ресурс за српски језик. Његов значај произлази из чињенице да омогућава формално представљање значења путем синсета, њихових дефиниција и међусобних семантичких односа, као и повезивање српског језичког материјала са широм мрежом међусобно повезаних ресурса типа ворднет посредством интерлингвалног индекса. На тај начин Српски ворднет чини основу за читав низ поступака у обради природног језика у којима је неопходно експлицитно моделовање значења, укључујући разрешење вишезначности, семантичко претраживање, анализу сентимента и друге облике аутоматског разумевања текста. У новијим описима ресурса наводи се да верзија коришћена у актуелним истраживањима садржи 25.320 синсета и 36.570 лема (Petalinkar 2025), док се у ширем прегледу језичких технологија за српски истиче да је последња јавно евидентирана верзија ресурса из 2018. године и да су неки домени у мрежи боље попуњени од других (Krstev and Stanković 2022, 2023). Ови подаци показују да Српски ворднет представља функционалан и истраживачки релевантан ресурс, али и да његова доступност, покривеност и даљи развој још увек не одговарају у потпуности потребама савремених система за ОПЈ.

Ограничења Српског ворднета односе се и на обим и на структуру самог ресурса. У поређењу са Принстонским ворднетом, српска мрежа је знатно мања, што непосредно утиче на покривеност лексике и значења. Поред тога, досадашњи радови указују на неуједначену попуњеност појединих делова мреже, као и на потребу да се, поред додавања нових синсета, систематски допуњују и постојећи синсети новим лексичким јединицама. То потврђују и раније анализе, према којима је просечан број лексема по синсету у српском ворднету мањи него у PWN-у, 1,66 према 2,01, што може да указује на недовољну попуњеност синсета. Део синсета представља значења специфична за српски, односно балкански језички простор и нема директан еквивалент у PWN-у (Stanković, Mladenović, et al. 2018). У истом контексту важно је имати у виду и то да део вредности придружених Српском ворднету није настао као резултат непосредне анотације за српски језик, већ је наслеђен путем поравнавања са спољним ресурсима. То се посебно односи на оцене сентимената преузете из SentiWordNet-а преко ILI механизма, што је омогућило иницијалну употребу ресурса у анализи сентимената, али је истовремено оставило отвореним питање њихове језичке и семантичке адекватности у српском контексту (Mladenović et al. 2015; Petalinkar 2025). Новија истраживања управо зато полазе од потребе да се такве наслеђене вредности не прихватају као коначне, већ да се ревидирају и поново процењују на основу података и метода прилагођених српском језику.

Последице оваквог стања непосредно се одражавају на примену ресурса у задацима аутоматског разумевања текста. Када је семантичка мрежа ограниченог обима, неравномерне покривености и делимично заснована на наслеђеним вредностима, а не на верификованим вредностима за српски језик, њена употреба у задацима који зависе од прецизног разликовања значења остаје методолошки и практично ограничена. То се посебно односи на разрешење вишезначности речи, али и на анализу сентимената, класификацију текста, семантичко повезивање и друге поступке у којима је потребно поуздано повезати језичку појаву у тексту са одговарајућим значењем у мрежи. Управо у том смислу Српски ворднет у овој дисертацији

није посматран само као постојећи лексички ресурс, већ као централни истраживачки случај на коме се јасно уочавају и потенцијал семантичких мрежа за српски језик и њихова садашња ограничења. Потреба за аутоматизованим проширивањем мреже, редакцијом наслеђених вредности и развојем поузданих метода за повезивање текста са синсетима не произлази, дакле, из општег интереса за унапређење ресурса, већ из конкретних ограничења која тренутно утичу на домет и поузданост поступака обраде природног језика за српски језик. Истовремено, та ограничења упућују и на шири истраживачки задатак ове дисертације, који не обухвата само унапређење самог ресурса, већ и развој метода његове примене у задацима аутоматског обележавања текста, повезивања значења и анализе сентимената. Овакво постављање проблема у складу је и са ширим истраживачким оквиром дисертације, у којем је Српски ворднет одређен као носећа студија случаја за испитивање могућности аутоматизације развоја и примене семантичких мрежа у српском језику.

2.13 Аутоматско и полуаутоматско проширивање ворднета у другим језицима

Аутоматско и полуаутоматско проширивање ворднета у другим језицима развијало се као одговор на ограничења ручне изградње оваквих ресурса. Тај поступак је по правилу дуготрајан, скуп и у великој мери зависан од доступности лексикографског и лингвистичког рада. У језицима са ограниченим ресурсима овај проблем је посебно изражен. Због тога су бројна истраживања била усмерена ка поступцима који омогућавају брже формирање почетне мреже, њено накнадно проширивање и постепену редакцију. У том оквиру развијени су различити приступи, од метода заснованих на превођењу и двојезичним речницима, преко поступака који укључују дистрибутивне и векторске репрезентације, до новијих приступа који се ослањају на велике језичке моделе.

У ранијој фази ових истраживања посебно су били заступљени приступи који полазе од већ постојећег ворднета (најчешће Принстонског) који је на енглеском језику, и настоје да његове синсете пренесу у циљни језик. У тим поступцима кандидати за еквиваленте у другом језику генеришу се аутоматски, на основу више извора, као што су двојезични речници, ресурси за машинско превођење и сродни лексички ресурси. Добијени кандидати се затим рангирају и филтрирају ради издвајања највероватнијих решења (Lam et al. 2014). Вредност ових приступа огледа се пре свега у томе што показују да се изградња почетне структуре ворднета може знатно убрзати. Истовремено, они јасно показују да коначна решења и даље зависе од ручне или полуаутоматске валидације.

Следећи корак у развоју ових метода представљају приступи у којима се преводни кандидати не прихватају само на основу формалног поклапања. Уместо тога, они се додатно проверавају на основу семантичке блискости. У ту сврху у поступак проширивања уводе се векторске репрезентације речи, које омогућавају да се кандидати рангирају и филтрирају у складу са дистрибутивном сличношћу²⁰. На тај начин смањује се број оних решења која су формално блиска изворној јединици, али семантички непоуздана. Овакви приступи јасно показују

²⁰ Дистрибутивна сличност представља степен сличности између речи утврђен на основу контекста у којима се оне јављају. Заснива се на претпоставци да речи које се појављују у сличним језичким окружењима најчешће имају сродна значења или сличне функције. У векторским моделима изражава се близином њихових векторских репрезентација.

ограничења чисто преводних метода и потребу да се у поступак укључи и додатни семантички критеријум (Al Tarouti and Kalita 2016).

У новијим истраживањима тежиште се постепено помера ка генеративним моделима. Уместо да се аутоматски поступак своди само на превођење појединачних лема, све чешће се испитује могућност да модел, на основу упутстава и контекста, предложи лексеме и структуре релевантне за циљни језик. Такав помак видљив је у радовима који велике језичке моделе користе за аутоматску изградњу двојезичних ворднета, односно за генерисање и процену кандидата унутар синсета (Bergh et al. 2025). Слично томе, велики језички модели почињу да се користе и у задацима анотације текста синсетима и проширивања синсета у језицима са скромнијом традицијом дигиталних лексичко-семантичких ресурса. Тиме се показује да они могу бити корисни у предлагању нових лексичких јединица, али и да њихови излази и даље захтевају пажљиву евалуацију (Santoro et al. 2025).

Из досадашњих истраживања везаних за друге језике, може се закључити да се најпоузданији резултати не добијају ослањањем на један изолован поступак. Напротив, најуспешнији приступи комбинују више техника. Преводни поступци омогућавају иницијално генерисање кандидата. Дистрибутивни и векторски модели уводе додатну семантичку контролу. Велики језички модели додају могућност аутоматског предлагања и формулисања лексичко-семантичког садржаја. Ипак, заједничко ограничење свих ових приступа је да они не укидају потребу за валидацијом. Она не нестаје, већ се само помера у другу фазу процеса. Управо због тога се у новијим истраживањима све чешће говори о аутоматском и полуаутоматском проширивању као о континууму, а не као о два потпуно одвојена модела рада.

У контексту ове дисертације искуства из других језика имају двострук значај. С једне стране, она показују да је аутоматско или полуаутоматско проширивање ворднета већ потврђен правац истраживања, а не изолован експериментални поступак. С друге стране, она показују и то да различите технике дају различите резултате, у зависности од језика, расположивих ресурса и начина евалуације²¹. Зато се у овој дисертацији аутоматско превођење, односно аутоматска адаптација и проширивање Српског ворднета не посматра као једно унапред дато техничко решење. Оно се посматра као истраживачки проблем у оквиру кога је потребно упоредити различите радне токове и различите моделе, како би се утврдило који приступ даје најпоузданије резултате за српски језик.

Поред проширивања самих ворднета, посебан истраживачки значај има и њихово повезивање са другим лексичко-семантичким ресурсима, што отвара питање поравнања ресурса као наредни проблем.

2.14 Поравнање лексичко-семантичких ресурса

Повезивање лексичко-семантичких ресурса представља једно од важних питања у савременој рачунарској лингвистици, нарочито у истраживањима која се односе на интеграцију различитих извора лексичког и семантичког знања. Под тим се, у најширем смислу, подразумева успостављање кореспонденције између јединица различитих ресурса, као што су синсети, дефиниције, речничке ставке или друге семантички релевантне јединице. Основна сврха таквог поступка јесте повећање интероперабилности између ресурса, проширење

²¹ у фази ручне допуне ворднета, синсетови су пружимани из енглеског и онда прилагођавани за српски: ни литерали нису директно превођени, ни глосе а посебно не примери употребе. Претходни покушаји аутоматског превођења нису били нарочито успешни и захтевали су значајне корекције.

покривености и омогућавање њихове заједничке употребе у различитим задацима обраде природног језика (Kwong 1998; Matuschek 2015).

Потреба за оваквим повезивањем уочена је још у ранијим радовима посвећеним ворднету и сродним лексичким ресурсима. У једном од старијих, али тематски и даље релевантних прилога, показано је да ворднет не треба посматрати као изолован ресурс, већ као чвориште које се може доводити у везу са другим лексикографским изворима ради обogaћивања и бољег структурирања знања. У том контексту је указано на значај комбиновања текстуалних описа и структурних информација при успостављању веза између различитих ресурса, чиме је рано постављена истраживачка основа за касније радове о поравнању значења (Kwong 1998).

Касније је овај проблем формализован знатно прецизније. У оквиру истраживања посвећеног поравнању значења лексичких ресурса представљеног у овој дисертацији, задатак је дефинисан као проналажење одговарајућих веза између значења која су описана у различитим ресурсима. Посебно је наглашено да се тај проблем не може свести на једноставно пресликавање један-према-један, јер различити ресурси често различито сегментирају простор значења, различито формулишу дефиниције и различито организују семантичке релације. Тиме је јасно показано да је поравнање значења²² пре свега проблем међусобног усклађивања хетерогених модела лексичко-семантичке репрезентације ради заједничког коришћења у задацима ОПЈ, а не само технички поступак повезивања идентификатора (Matuschek 2015).

У новијим истраживањима ова област је додатно систематизована. У систематском прегледу литературе посвећеном лексичком поравнању анализирани су различити поступци мерења и имплементације оваквог поравнања, уз указивање на чињеницу да су у досадашњим истраживањима коришћене веома разнорodne методе и критеријуми. Показано је да се приступи могу грубо груписати на методе засноване на текстуалној сличности, методе које користе структурне особине ресурса и хибридне моделе који комбинују више извора информација. Истовремено је указано и на то да још увек не постоји потпуно уједначен евалуациони оквир, што отежава непосредно поређење различитих решења и додатно потврђује да је реч о методолошки отвореној истраживачкој области (Srivastava et al. 2025).

За истраживање из ове дисертације посебно су важни радови у којима је проблем сужен на монолингвално поравнање значења у оквиру лексикографских извора. У раду у коме је овај задатак формулисан као класификациони проблем показано је да се поравнање може третирати као препознавање семантичког односа између парова значења из различитих монолингвалних ресурса. При томе су комбиноване ручно издвојене особине и технике учења (обучавања) репрезентација (енг. *representation learning*), а наглашено је да овакво повезивање омогућава проширење доменске покривености и интеграцију података из различитих извора. Овакав приступ је важан јер показује да се поравнање не мора посматрати само као низ хеуристичких правила, већ и као задатак погодан за надгледано машинско учење (Ahmadi and McCrae 2021).

У каснијем и обимнијем истраживању исти проблеми су додатно продубљени у правцу монолингвалног поравнања значења и дефиниција у лексикографским ресурсима. Посебно је истакнуто да се главне потешкоће јављају због разлика у грануларности значења, разлика у покривености и разлика у самом начину описивања значења у два ресурса. У том раду је уведен и богатији систем семантичких односа између јединица, који укључује потпуно поклапање, уже значење, шире значење, сродност и одсуство релевантне везе. На тај начин је показано да

²² Поравнање значења представља поступак успостављања семантички оправданих веза између значења забележених у различитим лексичким ресурсима. Оно може обухватити потпуно или делимично подударање значења и не своди се на формално упаривање њихових идентификатора.

однос између јединица различитих ресурса није нужно бинарног типа, већ да може имати више степена и више облика. За потребе евалуације развијен је и вишејезички скуп података са ручно означеним односима, што овом истраживању даје посебну методолошку тежину (Ahmadi 2022).

За тему ове дисертације наведени радови су нарочито значајни зато што омогућавају јасније формулисање потешкоћа које се јављају при повезивању српског ворднета са дефиницијама из описних речника. Најпре, уочава се неусклађеност у грануларности значења. Речничка организација је, по правилу, заснована на лексеми и њеним значењима, док је у ворднету основна јединица синсет, односно појединачно лексичко значење унутар мреже. Због тога једна речничка дефиниција може одговарати једном синсету, али може обухватити и шири или ужи семантички опсег од онога који је у ворднету експлицитно издвојен. Овај проблем је у литератури препознат као извор највећег броја грешака у поравнању (Ahmadi 2022; Matuschek 2015).

Даље, уочавају се значајне разлике у природи и функцији дефиниција. У описним речницима дефиниција је, по правилу, самостални елемент речничког чланка која треба да опише значење речи. У ворднету, међутим, глоса не функционише изоловано, већ као један од елемената у оквиру шире мреже значења и семантичких релација. То значи да се у ворднету значење не одређује само текстом дефиниције, већ и позицијом синсета унутар мреже. Последица тога је да површинска сличност између речничке дефиниције и глосе не мора нужно значити и потпуно поклапање значења, што је више пута истакнуто у досадашњим истраживањима (Ahmadi 2022; Ahmadi and McCrae 2021; Kwong 1998).

Поред тога, разликује се и сама структура семантичке организације ресурса. Електронски речници најчешће не располажу експлицитно формализованом мрежом односа каква постоји у ворднету, где су хиперонимија, хипонимија, меронимија и друге релације кључни део репрезентације. Због тога се при повезивању ова два типа ресурса не упоређују две истоветно организоване семантичке структуре, већ лексикографски ресурс заснован на дефиницијама и лексичко-семантичка мрежа заснована на синсетима и релацијама. Управо та структурна несиметрија представља једно од кључних методолошких ограничења у овом типу истраживања (Matuschek 2015; Srivastava et al. 2025).

Најзад, у условима рада са ресурсима који још увек нису у потпуности завршени, потребно је рачунати и са извесним степеном шума, неуједначености и непотпуности. То се може односити и на сам речнички материјал и на семантичку мрежу, нарочито у случајевима када поједина лексичка поља нису равномерно развијена, када глосе нису уједначене по дужини и квалитету или када постојећа покривеност није довољна. Овај аспект је у новијим радовима такође препознат као важан фактор који утиче на успех метода поравнања и на интерпретацију добијених резултата (Ahmadi 2022; Srivastava et al. 2025).

У ширем методолошком оквиру, проблем поравнања лексичко-семантичких ресурса све чешће се разматра и у контексту повезаних лингвистичких података и вишејезичне интероперабилности. Прегледни радови посвећени лингвистичким повезаним подацима показују да се повезивање ресурса не своди само на локално упаривање појединачних јединица, већ се посматра као део ширег поступка генерисања, међусобног повезивања, објављивања и валидације лингвистичких ресурса. Посебно је значајно то што се у овим радовима наглашава потреба за јасним моделима репрезентације, стабилним идентификаторима и стандардизованим односима између лексичких јединица, значења и

дефиниција, што је непосредно релевантно и за повезивање ворднета са електронским речницима (Gromann et al. 2024; Khan et al. 2022).

Посебан нови правац представљају истраживања у којима се велики генеративни језички модели користе као средство за избор значења, поређење дефиниција или генерисање проширених описа значења на основу пажљиво обликованих промптова. У тим радовима показано је да се задатак избора одговарајуће дефиниције може формулисати као интеракција модела са контекстом, значењима кандидата и експлицитно задатим ограничењима. Истовремено је показано и да на квалитет резултата значајно утичу структура промпта, распоред и баланс примера, као и начин формулисања дефиниција, што је посебно важно у условима вишејезичности и недовољно развијених језичких ресурса. Ови налази су непосредно релевантни за методологију ове дисертације, јер указују на то да се поравнање српског ворднета са речничким дефиницијама може испитивати не само класичним методама сличности и класификације, већ и контролисаном применом генеративних модела кроз промптове (Meconi et al. 2025; Sumanathilaka et al. 2025).

Полазећи од наведених увида, може се закључити да повезивање српског ворднета са дефиницијама из електронских речника не представља једноставан поступак проналажења еквивалентних јединица, већ сложен задатак у којем се истовремено јављају проблеми различите грануларности, различите функције дефиниција, структурне несиметрије ресурса и извесног степена непотпуности. Управо због тога се у овој дисертацији овај проблем не посматра као споредни технички корак, већ као засебан лексичко-семантички задатак, релевантан и за развој ресурса и за њихову каснију примену у аутоматском разумевању текста.

3 Методологија

Методологија истраживања организована је тако да прати ток од развоја и провере појединачних слојева Српског ворднета ка њиховој примени у задацима аутоматског разумевања текста. Најпре се описује поступак одређивања и накнадног прилагођавања оцена сентимента синсета у Ворднету, заснован на структурним односима између синсета, њиховим глосама и моделима машинског учења. Затим се приказује формирање посебног скупа за евалуацију оцена сентимента помоћу генеративног модела, који треба да послужи као контролни механизам за проверу случајева у којима постојеће пресликавање из *SentiWordNet*-а није довољно поуздано за српски језик.

Другу методолошку целину чине поступци за разрешење вишезначности речи. Прво се описује приступ заснован на генеративном моделу, у којем се за дату лексему у контексту бира најприкладнији синсет из унапред припремљеног инвентара значења, а затим поступак дестилације, којим се знање добијено сложенијом анотацијом преноси на мањи и рачунски ефикаснији модел. Након тога следе методе усмерене на проширивање и повезивање лексичко-семантичких ресурса: превод синсета, као поступак допуне језички специфичних елемената ресурса, и поравнање Српског ворднета са другим лексичко-семантичким изворима уз употребу генеративних модела. Завршни део методологије посвећен је провери употребљивости тако обогаченог ресурса у низводним задацима, најпре кроз екстерну евалуацију у анализи сентимента, а затим кроз циљану проверу утицаја слоја за разрешење вишезначности речи на избор преводних еквивалената у машинском превођењу.

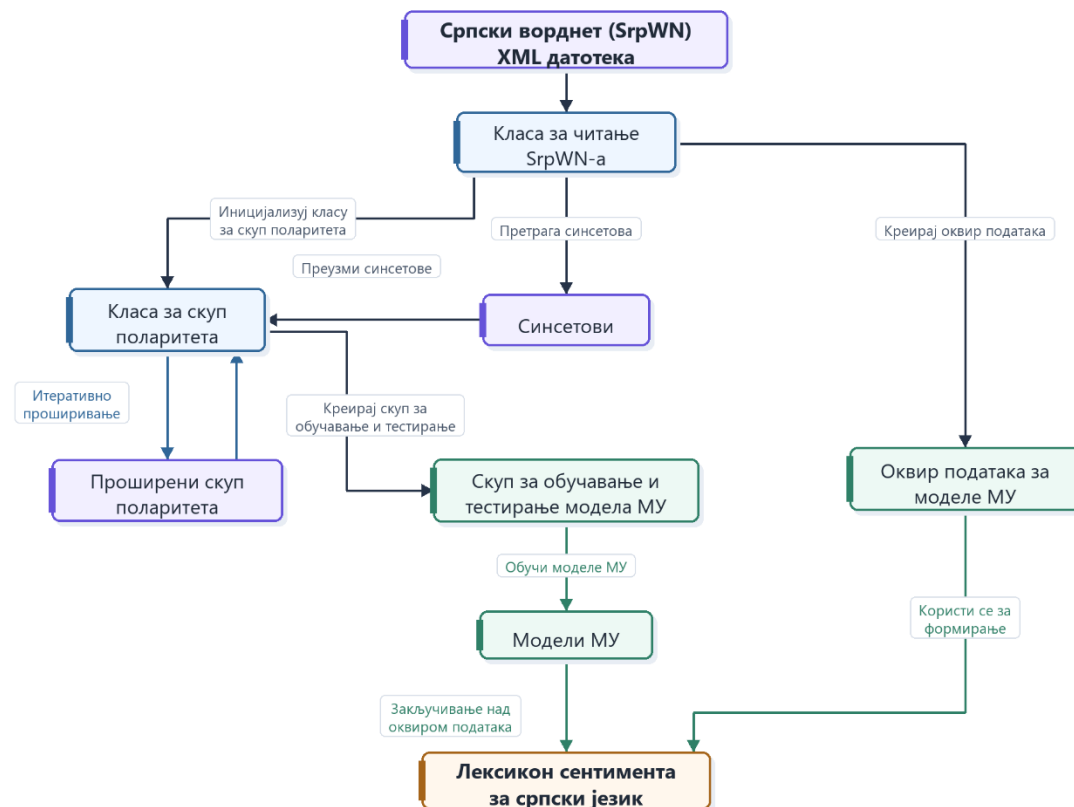
3.1 Оцена сентимента у ворднету

У овом одељку описане су методе које су примењене да би се анотација сентимента из SWN-а прилагодила постојећем Српском ворднету, са циљем њиховог бољег усклађивања са језичким специфичностима српског језика (Petalinkar et al. 2025). Полазна тачка радног тока јесте XML репрезентација целокупног или делимичног Српског ворднета, док је крајњи резултат лексикон сентимента који обухвата све синсете из обрађене репрезентације.

У ту сврху развијен је софтверски модул за учитавање и програмску обраду XML структуре ворднета. Овај модул омогућава систематско претраживање синсета према различитим критеријумима, као и њихово организовање у скупове у складу са додељеним или израчунатим поларитетом сентимента. Тако формиран скупови могу се користити као:

1. скупови за обучавање модела машинског учења,
2. полазни скупови за итеративно проширење путем релација дефинисаних у XML структури,
3. основа за примену модела над глосама синсета ради одређивања вредности сентимента.

Посебна функционалност модула односи се на издвајање и структурирање глоса синсета, које се затим користе као текстуална репрезентација значења у поступцима машинског учења и каснијој процени поларитета. Целокупан поступак визуелно је приказан на Слици 3.1.



Слика 3.1 Радни ток формирања лексикона сентимента за српски језик на основу Српског ворднета, проширивања скупа поларитета и обучавања модела машинског учења.

За добијање коначних оцена сентимента примењене су две варијанте агрегирања резултата класификатора. Прва варијанта одговара поступку описаном у оригиналном раду (Esuli and Sebastiani 2006), као што је представљено у одељку 2.6. Формирано је осам тернарних класификатора, добијених комбинацијом различитих вредности параметра K (0, 2, 4 и 6 корака експанзије) и различитих алгоритама учења. Сваки класификатор додељује синсету бинарну ознаку²³ припадности одређеној класи (позитивној или негативној), при чему се вредност кодира као 0 или 1. Коначна оцена сентимента добија се израчунавањем просечне вредности преко свих осам класификатора. На тај начин добијена скала обухвата осам дискретних вредности у интервалу $[0, 1]$, са увећањем од $1/8$ (0,125)²⁴. Друга варијанта заснива се на коришћењу излазних вероватноћа класификатора, односно процене вероватноће да синсет припада одређеној класи. Уместо бинарне одлуке (0 или 1), у овом приступу се агрегирају континуалне вероватносне вредности, чиме се добија финија и потенцијално информативнија скала оријентације сентимента.

Сви алати развијени у оквиру овог истраживања имплементирани су у програмском језику Python, који представља најчешће коришћено окружење за истраживања у области обраде природног језика и машинског учења. Избор овог језика мотивисан је његовом зрелом мрежом међусобно повезаних ресурса типа ворднет, библиотека и широком применом у академским истраживањима. За рад са лексичким ресурсима коришћена је библиотека NLTK, која

²³ Тернарни класификатор има три могућа исхода: позитиван, негативан и објективан. За потребе усредњавања, за сваку класу засебно бележи се 1 ако је синсет сврстан у њу, а 0 у супротном. Просек ових вредности представља удео класификатора који су синсет сврстали у дату класу.

²⁴ У обе варијанте крајњи резултат чине засебне оцене позитивности и негативности синсета у интервалу $[0, 1]$; у овој фази оне се не свде на једну коначну класу.

обезбеђује интерфејсе за PWN и SWN. На основу ње развијен је модул за учитавање и обраду XML репрезентације Српског ворднета.

За обучавање модела на скуповима формираним у овом истраживању коришћене су постојеће имплементације класификационих алгоритама из библиотеке *scikit-learn*, која је коришћена и за њихову евалуацију. Из методологије изградње SWN-а преузет је општи поступак проширивања скупова за обучавање, класификације глоса и агрегирања резултата више класификатора (Esuli and Sebastiani 2006). За експерименте са моделима заснованим на неуронским мрежама коришћени су TensorFlow и Keras, док је библиотека transformers (Hugging Face) коришћена за рад са трансформерским архитектурама и фино подешеним језичким моделима. За обраду и анализу података коришћене су библиотеке pandas и numpy, док је matplotlib употребљен за визуализацију резултата.

Припрема Српског ворднета за рад у Python окружењу подразумевала је израду програмског интерфејса компатибилног са омотачем PWN-а у оквиру NLTK модула. Иако већ постоје решења за интеграцију ворднета заснованих на XML-у, као што су Eurown за EuroWordNet (Kahusk, n.d.), по чијој је шеми Српски ворднет и конструисан, и ruwnxml за мађарски ворднет у оквиру BalkaNet-a (Indig and Miháltz, n.d.), њихова примарна намена усмерена је на приступ и претраживање података.

Показало се, међутим, да овај иницијални скуп није довољан за стабилну обуку модела. Стога је извршено постепено проширење иницијалних група укључивањем синсета чије су вредности сентимента већ биле доступне путем пресликавања из енглеске верзије SWN-а, као и из ручно означеног скупа Micro-WN(Op)-3.0. Посебно су разматрани синсети код којих је разлика између POS и NEG вредности износила најмање 0,5 у корист доминантне оријентације. У анализираној верзији Српског ворднета 160 синсета испуњавало је услов за позитивну, а 279 за негативну оријентацију.

Иако је овај критеријум формално јасан, показало се да директно ослањање на пресликане вредности није довољно. Због тога је сваки кандидат за проширење скупа подвргнут ручној провери. У том поступку уочено је да одређени синсети, иако формално означени као позитивни или негативни у изворном ресурсу, у српском језику не носе одговарајућу вредносну оријентацију. Тако је, на пример, из позитивне групе уклоњен синсет ‘ENG30-00971650-v; оцијенити, оквалификовати, прогласити, изрећи суд о некоме’, будући да радња вредновања сама по себи не подразумева позитиван став. Слично томе, синсет ‘ENG30-09312843-n; инфективни агенс’ представља неутралан научни појам и не може се сматрати ни позитивним ни негативним.

Из негативне групе искључени су, између осталих, синсет ‘ENG30-10940669-n; Достојевски’, који као сопствено име нема инхерентну негативну оријентацију, као и синсет ‘ENG30-00476819-a; удобан’, које је као резултат пресликавања погрешно негативно оцењен, јер у српском језику овај појам има јасно позитивну конотацију. Ови примери показују да механичко преношење вредности из енглеског ресурса може довести до семантички неприхватљивих резултата, те да је ручна корекција неопходан корак у процесу адаптације.

Након ручног филтрирања, у позитивни скуп укључен је 121 синсет, а у негативни 192 синсета из проширених група. За формирање објективног скупа (SSo) примењен је другачији поступак. Полазиште је било директно пресликавање објективних синсета из Micro-WN(Op)-3.0, након чега су додати сви синсети у Српском ворднету код којих су POS и NEG вредности биле једнаке нули. Из овог скупа накнадно су уклоњени синсети који садрже лексеме присутне у српском лексикону сентимента Senti-Pol-sr (Stanković et al. 2022), као и они код којих је ручном

анализом утврђено да ипак носе вредносну оријентацију (нпр. синсет *издаја*). Коначни иницијални скуп за обуку обухватао је 149 позитивно означених, 219 негативно означених и 19.475 објективних синсета. Већи број негативних синсета одражава структуру група кандидата, док је доминација објективних синсета очекивана, будући да већина лексема не носи експлицитну вредносну оријентацију.

Итеративни поступак проширења скупова сентимента имплементиран је у оквиру класе *PolaritySets*, која представља програмску апстракцију позитивног (SSp), негативног (SSn) и објективног (SSo) скупа синсета. Основна сврха ове класе није била само техничка организација података, већ омогућавање контролисаног ширења поларитетних категорија на основу структурних релација унутар Српског ворднета.

Полазиште процеса представља иницијални скуп ручно одабраних синсета. У нултом кораку (T_0) позитивни и негативни скупови садрже искључиво проверене синсете јасне оријентације сентимента. У наредним корацима примењује се експанзија заснована на релацијама које или очувавају или инвертују поларитет. У свакој итерацији, за сваки синсет из позитивног скупа, проналазе се синсети повезани релацијама које се могу тумачити као поларитетно конзистентне (нпр. *similarity*, *derived-from*, *attribute*, *also-see*). Ови синсети се додају позитивном скупу, под условом да већ нису експлицитно означени као негативни. Аналоган поступак примењује се на негативни скуп. Истовремено, релација директне антонимије тумачи се као оператор инверзије поларитета: антоними позитивних синсета додају се негативном скупу, и обрнуто.

Кључна карактеристика имплементације јесте одржавање међусобне искључивости позитивног и негативног скупа. Уколико синсет током итерације испуни услове за обе категорије, он се уклања из оба скупа и, уколико није експлицитно означен као субјективан (позитиван или негативан), остаје у објективном скупу. На тај начин се спречава акумулација контрадикторних ознака и ограничава ширење шума у подацима за обучавање.

Објективни скуп (SSo) третира се комплементарно: сваки синсет који уђе у позитивни или негативни скуп аутоматски се уклања из објективног. Оваква организација обезбеђује да се категорија објективности дефинише као резидуална класа, а не као скуп који се може независно проширивати. Ово је у складу са концептуализацијом објективности као одсуства позитивне и негативне оријентације, а не као засебне семантичке особине.

Итеративна природа поступка омогућава постепено повећање покривености, али истовремено уводи ризик од ширења семантичког шума. Са сваким наредним кораком повећава се број синсета који су индиректно повезани са иницијалним језгром, што повећава одзив (енгл. *recall*), али потенцијално смањује прецизност (енгл. *precision*). Анализом је утврђено да након шест итерација долази до значајног раста скупа уз релативно мали добитак у квалитету, што је условило избор скупова T_0 , T_2 , T_4 и T_6 као репрезентативних конфигурација.

Поред проширења путем структурних релација, имплементација је омогућила и екстракцију глоса синсета и њихову трансформацију у табеларни формат погодан за машинско учење. За сваки синсет бележи се текст глосе и бинарне ознаке припадности позитивној и негативној класи. На овај начин се лексичко-семантичка структура ворднета трансформише у скуп примера погодан за надгледану класификацију, чиме се успоставља директна веза између симболичког ресурса и статистичког модела.

Методолошки гледано, ова имплементација представља компромис између чисто лексикографског приступа и потпуно статистичког моделирања. Полазна семантичка

структура ворднета служи као механизам контролисаног проширења, док машинско учење омогућава да се из глоса синсета извуче општи образац оријентације сентимента. Оваква хибридна архитектура омогућава већу флексибилност и прилагодљивост српском језичком систему него што би то било могуће једноставним пресликавањем вредности из енглеског ресурса.

Табеле, формиране у овом одељку након итеративне експанзије иницијалних скупова синсета, представљају основу за обуку модела за аутоматску класификацију оријентације сентимента. Нека су:

- SS_p^k — скуп синсета означених као позитивни након k корака експанзије,
- SS_n^k — скуп синсета означених као негативни након k корака експанзије,
- SS_o^k — скуп синсета означених као објективни.

За сваку вредност k дефинише се скуп за обучавање:

$$T_k = SS_p^k \cup SS_n^k \cup SS_o^k$$

Класификација сентимента реализована је путем два бинарна класификатора, у складу са оригиналним приступом изградње SWN-а. Дефинишу се функције:

где је:

- $P_k(s) = 1$ ако синсет припада позитивној класи,
- $N_k(s) = 1$ ако синсет припада негативној класи.

Класификатор P_k обучаван је над скупом:

$$\mathcal{D}_P^k = SS_p^k \cup (SS_n^k \cup SS_o^k)$$

при чему су позитивни примери елементи SS_p^k , а негативни примери унија $SS_n^k \cup SS_o^k$. Аналогно томе, класификатор N_k обучаван је над:

$$\mathcal{D}_N^k = SS_n^k \cup (SS_p^k \cup SS_o^k)$$

Тако се проблем тернарне класификације редукује на две бинарне одлуке.

Коначна одлука Φ_k дефинисана је као:

$$\Phi_k(s) = \begin{cases} POS, & \text{ако } P_k(s) = 1 \wedge N_k(s) = 0 \\ NEG, & \text{ако } P_k(s) = 0 \wedge N_k(s) = 1 \\ OBJ, & \text{у супротном} \end{cases}$$

Ова формулација омогућава да се синсет означи као објективан у случају конфликтних одлука или одсуства обе позитивне одлуке, што је у складу са концептуализацијом објективности као комплементарне категорије.

Нека је $g(s)$ глоса синсета s . За свако $s \in T_k$ конструисана је векторска репрезентација:

$$\mathbf{x}_s = f(g(s))$$

где функција f зависи од примењеног приступа кодирању текста.

Први приступ заснива се на моделу вреће токена, при чему се текст глосе представља као вектор фреквенција речи. Над овом репрезентацијом примењена је TF-IDF трансформација, са косинусном нормализацијом. Овај приступ директно прати методологију оригиналног SWN-a и омогућава упоредивост резултата.

Други приступ подразумева примену компактних неуронских модела реализованих у оквиру *TensorFlow/Keras* окружења. Глосе су кодиране употребом слоја *TextVectorization*, након чега су испитиване архитектуре засноване на рекурентним неуронским мрежама и лаганим трансформерским структурама. Ови модели омогућавају експлицитно моделирање редоследа речи и контекстуалних зависности.

Трећи приступ заснива се на фино подешеним трансформерским моделима претходно обученим на великим корпусима. За потребе класификације коришћена је класа *AutoModelForSequenceClassification*, при чему је на основни језички модел додат класификациони слој за бинарну одлуку. Модели су фино подешавани над скуповима за тренинг, уз примену механизма раног заустављања ради спречавања прекомерног прилагођавања (енгл. *overfitting*).

За сваку вредност $k \in \{0,2,4,6\}$ обучавани су засебни модели Φ_k . На тај начин формиран је комбинација класификатора, аналогно оригиналном SWN приступу. Коначна оцена сентимента синсета може се добити агрегирањем резултата више модела, чиме се умањује варијабилност појединачних конфигурација скупова за обучавање.

Будући да су класе биле неуравнотежене (посебно је доминантна објективна класа), као примарна метрика за избор модела коришћена је F-мера (енгл. *F-score*, *F-measure*), која представља хармонијску средину прецизности (енгл. *precision*) и одзива (енгл. *recall*). Поред F-мере, анализиране су и појединачне вредности прецизности, одзива и укупне тачности (енгл. *accuracy*), али је избор најбољег модела заснован на F-мери, с обзиром на осетљивост тачности на дисбаланс класа.

На основу обучених модела, креирани су нови лексикони сентимента применом добијених класификатора на глосе свих синсета у Српском ворднету. Полазна тачка била је постојећа структура синсета, док су вредности сентимента израчунаване искључиво на основу текстуалне репрезентације глосе, без директног ослањања на иницијално пресликавање из енглеског SWN-a.

Конструисано је седам нових лексикона, означених као S_1 – S_7 . Ради прегледности, постојеће директно пресликавање вредности сентимента из SWN-a на Српски ворднет означено је као лексикон S_0 . Лексикони S_1 и S_2 заснивају се на класичним методама машинског учења, S_3 и S_4 на компактним неуронским архитектурама, док су S_5 , S_6 и S_7 добијени применом фино подешених великих језичких модела.

У случају лексикона S_1 , примењени су SVM и *Bernoulli Naive Bayes* класификатори у оквиру ансамбла модела обучених над скуповима $\overset{\text{OBL}}{T}_0, T_2, T_4, \overset{\text{OBL}}{T}_6$ израчунате су вредности:

$$POS(s) = P_k(s) \cdot (1 - N_k(s))$$

$$NEG(s) = N_k(s) \cdot (1 - P_k(s))$$

$$OBJ(s) = 1 - POS(s) - NEG(s)$$

Ове формуле директно одражавају правило комбиновања бинарних одлука, али омогућавају континуалну интерпретацију резултата у случају када класификатори враћају вероватноће припадности класи. Коначне вредности у лексикону S_1 добијене су усредњавањем резултата свих модела у оквиру ансамбла.

Лексикон S_2 направљен је на идентичан начин, при чему су уместо SVM-а коришћени *AdaBoost* и *Bernoulli Naive Bayes* класификатори. На тај начин омогућено је испитивање утицаја различитих класичних алгоритама машинског учења на стабилност и квалитет добијених сентимент оцена.

Лексикон S_3 заснива се на рекурентним неуронским мрежама, док је S_4 добијен применом компактне трансформерске архитектуре реализоване у Keras окружењу. У оба случаја класификатори су враћали вероватноћу припадности позитивној, односно негативној класи. Те вероватноће су коришћене у истим формулама за израчунавање *POS*, *NEG* и *OBJ*, а финалне вредности су добијене као средња вредност резултата модела обучених над различитим скуповима T_k .

Лексикони S_5 , S_6 и S_7 генерисани су применом фино подешених великих језичких модела — редом BERTiс, GPT2-Orao и Jerteh-355. Сваки од ових модела реализован је као *sequence classification* архитектура, која враћа вероватноћу да дата глоса припада одређеној класи. И у овом случају излазне вероватноће су трансформисане у оцене сентимента применом истих формула, чиме је обезбеђена методолошка упоредивост свих лексикона.

На тај начин добијен је скуп лексикона који се разликују по методолошком оквиру (класични модели, неуронске мреже, велики језички модели), али су конструисани у јединственом формалном систему израчунавања вредности сентимента. Оваква поставка омогућава директно поређење резултата и анализу утицаја различитих приступа на структуру и стабилност репрезентације сентимента у Српском ворднету. Сви генерисани лексикони, као и пратећи програмски код, јавно су доступни у репозиторијуму наведеном у прилогу дисертације, што омогућава **понављање** експеримената и поређење резултата као и обављање других експеримената.

3.2 Скуп за евалуацију оцена сентимента креиран генеративним моделом

У овом истраживању лексикон *Senti-pol-sr* коришћен је као референтни извор за одабир узорка погодног за анотацију помоћу BJM (Stanković et al. 2022). Конкретно, идентификовани су сви синсети из Српског ворднета који садрже литерале присутне у лексикону *Senti-pol-sr*, а који су истовремено имали неутралну вредност сентимента (0,0) у мапирању из SWN. На тај начин формиран је скуп кандидата код којих постоји потенцијално неслагање између постојећег поларитетског означавања и лексичке индикације поларитета у српском језику.

Идентификована су четири главна разлога за неслагања између вредности сентимента у српском језику и оних изведених из SWN-а. Прво, иако одређена реч може носити поларизујући сентимент, конкретно значење у којем се појављује не мора нужно имати такву конотацију. На пример, синсет ENG30-06828389-n (деф. „звездасти знак * који се користи у штампању“) у Српском ворднету садржи литерал „звезда“, који у другим синсетима може

имати позитивне конотације. Друго, вредности сентимента у SWN-у, које су генерисане методама машинског учења, могу бити нетачне. На пример, синсет ENG30-02585489-v (прогонити, угњетавати деф. „изазвати патњу“) има очигледно изражен негативан поларитет, али то није адекватно обележено ни у SWN-у па према томе ни у мапираним вредностима. Треће, уколико синсет није поравнат ни са једним синсетом у PWN-у, као у случају BILI-00000941 (нарицати, деф. “гласно изражавати жалост, запевати, тужити”), по правилу му се додељује неутрална вредност поларитета (0,0). Најзанимљивија хипотеза подразумева могућност да се одређено значење у енглеском језику третира као објективно, док у српском језику носи изражен сентимент. Међутим, у оквиру досадашње анализе такви синсети нису идентификовани.

Почетна анализа обухватила је 2.956 синсета од укупно 25.320 у оквиру Српског ворднета који су садржали литерале означене јасним поларитетом у лексикону *Senti-pol-sr*, а истовремено имали вредност поларитета (0,0) у мапирању из SentiWordNet-a (SWN). Од тог броја, 1.511 синсета повезано је са позитивним, а 1.445 са негативним сентиментом. Имајући у виду обим овог скупа, обрада свих синсета помоћу великих језичких модела оцењена је као непрактична јер је временски и финансијски захтевна. Стога је за даљу анализу насумично изабран узорак од 500 синсета.

На основу претходно дефинисаног узорка синсета, примењен је промпт конципиран за анализу сентимента помоћу ВЈМ, конкретно Mistral 7B – Instruct. Применом овог промпта (в. Слика 3.2) синсети су класификовани у четири категорије: позитивни, негативни, објективни и неправилно означени. Почетна класификација скупа од 500 синсета је 290 синсета означила као објективне, 102 као негативне, 33 као позитивне и 75 као неправилно категорисане.

Шаблон упита за почетну класификацију сентимента синсета

Mistral 7B – Instruct · учење без примера (zero-shot)

● УПИТ (ИЗВОРНИ ТЕКСТ) улаз: глоса синсета

Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment.

Sentiment treba da bude striktno klasifikovan kao "pozitivan", "negativan", ili "objektivan".

Nijedan drugi odgovor neće biti prihvaćen.

Tekst: {text}

Sentiment:

ОЧЕКИВАНИ ИЗЛАЗ МОДЕЛА: ЈЕДНА ОЗНАКА

● **позитиван**

● **негативан**

● **објективан**

● **други облик**
→ неправилан излаз

Напомена: „неправилно означени“ није четврта сентимент класа, већ категорија излаза који не одговара задатом формату.

Слика 3.2 Промпт²⁵ конципиран за анализу сентимента помоћу ВЈМ

²⁵ Писмо упита прилагођавано је писму обрађиваног текста, па су у различитим експериментима коришћени упити на ћирици или латиници. У дисертацији је при приказивању упита задржано писмо у којем су употребљени у одговарајућој програмској реализацији.

Промпт је конципиран у складу са принципима контролисане интеракције са ВЈМ моделима и обухватао је три кључне компоненте:

- 1) **Дефинисање улоге модела.** Моделу је експлицитно додељена улога експерта за анализу сентимента. Овим приступом настојало се да се активира аналитички режим одговора, усмерен на систематичну процену поларитета дефиниција синсета.
- 2) **Прецизне инструкције.** Промпт је јасно дефинисао задатак, захтевајући идентификацију и процену сентимента израженог у дефиницији синсета. Формулација инструкција смањивала је на најмању могућу меру могућност двосмислене интерпретације.
- 3) **Стандарди излаза.** Промптом је унапред одређен формат одговора, било у виду трочлане класификације (позитиван, негативан, неутралан), било у облику степеноване процене поларитета. Ово је омогућило конзистентну и упоредиву обраду резултата.

Промпт је финализован након експерименталног тестирања на мањем подскупу синсета. Упоредна анализа инструкција формулисаних на енглеском и српском језику показала је да промптови на српском језику дају стабилније и прецизније резултате. Максималан број токена одговора у излазу ограничен је на пет, како би се обезбедила контролисана и концизна класификација.

Анализа излаза указала је на постојање дупликата, будући да поједини синсети садрже више литерала присутних у лексикону *Senti-pol-sr*. На пример, синсет ENG30-01375831-a („славан, познат, чувен, значајан, реномиран, признат, прослављен: деф. опште прихваћен и уважен“) појављивао се више пута услед вишеструке лексичке покривености.

Након уклањања дупликата, коначни број синсета по категоријама износио је:

- 279 објективних,
- 97 негативних,
- 27 позитивних.

Категорија неправилно означених синсета није поновно квантитативно анализирана, јер уклањање дупликата у том сегменту није имало методолошки значај.

Ради детаљније квалитативне анализе, из сваке од три главне категорије добијене у претходној фази насумично је изабрано по 25 синсета. На тај начин формиран је уравнотежени узорак од 75 синсета, који омогућава контролисано поређење позитивних, негативних и објективних случајева, без доминације најзаступљеније класе. Овај узорак није коришћен за увођење новог балансираног скорa, већ као ограничен и проверљив скуп за испитивање поступка доделе вредности сентимента на нивоу синсета.

Овај уравнотежени узорак обрађиван је другим скупом промптова, који је заснован на моделу *mistral-small3.2*, покренутом локално преко система *Ollama*. Основна организација промптова задржала је елементе из претходне фазе: моделу је дата улога стручњака за анализу сентимента, задатак је формулисан као процена поларитета дефиниције синсета, а излаз је био ограничен на унапред задати скуп одговора. Међутим, циљ другог скупа промптова више није био разврставање синсета у шире категорије, већ добијање вредности компатибилних са формализмом *SentiWordNet-a*. Због тога су за сваки синсет независно процењиване две компоненте сентимента: позитивна компонента POS(s) и негативна компонента NEG(s).

За сваки синсет коришћена су два одвојена промпта: један за процену позитивног поларитета, а други за процену негативног поларитета. Такво раздвајање одговара основној претпоставци SentiWordNet-а да позитивни и негативни поларитет нису нужно међусобно искључиве категорије, већ две независне компоненте исте репрезентације сентимента. У позитивном промπτу модел је бирао једну од следећих категорија: „није позитиван“, „слабо позитиван“, „позитиван“, „веома позитиван“ и „екстремно позитиван“. У негативном упиту коришћене су аналогне категорије: „није негативан“, „слабо негативан“, „негативан“, „веома негативан“ и „екстремно негативан“.

Ради стабилизације излаза примењен је режим са малим бројем примера (енгл. *few-shot*), при чему су примери имали функцију граничних калибрационих тачака. Као горња граница позитивног поларитета коришћен је синсет ENG30-01123148-а, са српским литеаралима „частан“, „добар“ и „поуздан“. Овај пример је у промπτу означен као „екстремно позитиван“ и „није негативан“. Као доња граница негативног поларитета коришћен је синсет ENG30-01125429-а, са литеаралима „лош“ и „покварен“. Он је означен као „није позитиван“ и „екстремно негативан“. Као неутрални, односно објективни калибрациони пример коришћен је синсет ENG30-01818992-а, са литеаралом „неутралан“, означен као „није позитиван“ и „није негативан“. Ови примери нису третирани као обични примери емоционалног интензитета, већ као калибрационе тачке за пресликавање категоријских²⁶ одговора на скалу SentiWordNet-а.

У сам промпт је додатно унета напомена да се категорије „екстремно позитиван“ и „екстремно негативан“ не односе нужно на драмски или емоционални интензитет у свакодневном језику, већ на крајње вредности позитивног и негативног поларитета у SentiWordNet-у. Тиме је модел усмерен да процењује семантички поларитет појма описаног дефиницијом, а не само афективну јачину текста.

Посебна пажња посвећена је контроли формата излаза. Од модела је захтевано да за сваку процену врати структуриран одговор са два елемента: изабраном категоријалном оценом и кратким образложењем избора. У техничкој реализацији ова два елемента била су представљена JSON пољима *value* и *explanation*, што је омогућило програмску проверу формата и аутоматску обраду резултата. Поред тога што је JSON шема наведена у самом промπτу, коришћена је и програмска валидација структурираног излаза помоћу *Pydantic*²⁷ шеме у оквиру *LangChain/Ollama*²⁸ интерфејса. На тај начин су одговори модела аутоматски проверавани, а одговоре ван дозвољеног скупа или у неисправном JSON формату је било могуће одбацити или поново затражити.

Након добијања категоријских оцена, оне су пресликане у нумеричке вредности компатибилне са SentiWordNet формализмом. Пресликавање је изведено линеарно, са кораком 0,25: категорије „није позитиван“ и „није негативан“ одговарају вредности 0, „слабо позитиван“ и „слабо негативан“ вредности 0,25, „позитиван“ и „негативан“ вредности 0,5, „веома

²⁶ Категоријални одговори представљају оцене изражене избором једне од унапред дефинисаних, дискретних категорија, за разлику од оцена исказаних бројчаном вредношћу на континуираној скали.

²⁷ *Pydantic* је Python библиотека за валидацију података, проверу типова и серијализацију, заснована на Python *type hints* механизму. Омогућава да се подаци (нпр. из JSON-а, API захтева, конфигурационих датотека или LLM излаза) аутоматски провере и конвертују у структуре дефинисане класама.

²⁸ *LangChain/Ollama* интерфејс означава интеграцију између: *LangChain*-а – оквира за развој апликација заснованих на великим језичким моделима, и *Ollama*-е – платформе за локално покретање LLM модела (*Llama*, *Mistral*, *Gemma*, *Qwen* и др.). Практично, *LangChain* обезбеђује унификован API за рад са моделима, док *Ollama* локално извршава модел.

позитиван“ и „веома негативан“ вредности 0,75, а „екстремно позитиван“ и „екстремно негативан“ вредности 1,0. На тај начин су за сваки синсет изведене вредности POS(s) и NEG(s). Пропорционална нормализација је била предвиђена за случајеве да је збир POS(s) + NEG(s) већи од 1, али у обрађеном узорку такви случајеви нису забележени.

Добијене категоријалне оцене и њихове нумеричке вредности затим су коришћене као основа за анализу резултата, која је приказана у одељку 5.3.2. Поред коначних вредности POS и NEG, сачувани су и изворни категоријални одговори модела, како би се омогућила накнадна ручна провера спорних случајева.

3.3 Разрешење вишезначности речи – генеративни модел

У овом истраживању разрешење вишезначности речи постављено је као поступак избора најприкладнијег значења за дату лексему у конкретном контексту, при чему се избор врши у односу на унапред дефинисан инвентар значења. Полази се од претпоставке да су текстови претходно сегментирани на реченице, токенизовани и морфосинтаксички обележени, тако да су за сваку појаву речи бар доступне информације о лема (речнички облик) и врсти речи. За српски језик, као морфолошки богат језик, овај корак је посебно важан, јер једно исто значење може бити реализовано кроз већи број површинских облика, па лема и врста речи представљају основну везу између конкретног токена и релевантних значења из ресурса.

У оквиру експерименталне евалуације коришћено је више генерација великих језичких модела, као и три базна модела заснована на векторским репрезентацијама. Сви модели добијали су исти тип улаза: контекст реченице, циљну јединицу, лему, врсту речи и листу значења кандидата из инвентара значења. Поступак је спроведен у две итерације, при чему је у првој коришћена почетна, а у другој ревидирана (допуњена и коригована) верзија инвентара значења. Као основна мера успешности коришћена је тачност, дефинисана као однос исправно додељених значења и укупног броја аотираних јединица.

По потреби се у поступак укључују и додатни аотациони слојеви²⁹, нарочито слојеви за вишечлане лексеме и именоване ентитете. Због тога се у практичном току обраде најпре разматрају вишечлане јединице, а затим појединачни токени који носе значење, у које спадају именице, глаголи, придеви и прилози. За сваки пример, на основу леме и врсте речи, из инвентара значења издваја се скуп могућих значења кандидата. Тиме се задатак не поставља као слободно одређивање значења, већ као контролисан избор унутар ограниченог скупа могућих решења.

Овакво сужење простора одлучивања има више предности. Прво, обезбеђује се упоредивост аотације, јер су сви избори везани за исти инвентар значења. Друго, смањује се могућност да модел произведе значење које није поткрепљено ресурсом. Треће, добијени резултати могу се непосредно укључити у даљи ток аотације, провере и редакције ресурса.

Средишњи део поступка чини осмишљавање промпта. Промпт је обликован тако да модел добије јасно дефинисан задатак, ограничен скуп кандидата значења и строго задату структуру одговора. У општем облику, промпт садржи реченицу са означеном циљном јединицом, лему циљне јединице и нумерисану листу кандидата значења. Свако значење које је кандидат

²⁹ За једночлане јединице коришћене су лема и врста речи из једног аотационог слоја, док су за вишечлане јединице коришћени тип и лема вишечлане јединице из другог слоја. Иако су ове информације у аотацији структурно повезане, у извезеној TSV датотеци налазе се на различитим местима, па је код морао одвојено да их обради, а затим обједини

садржи идентификатор и глосу на српском језику. Оваква организација упита омогућава да модел избор врши над јасно представљеним и међусобно разграниченим кандидатима, а не над неограниченим простором могућих тумачења.

Ради јаснијег увида у структуру поступка, корисно је навести један стварни, скраћени пример. У једном од примера из корпуса који је обрађиван циљна реч је глагол **одржати** у реченици:

*У петој сезони програм је ****одржан**** у складу са новим концептом: један млади фудбалер из сваке земље изабран је као њен представник.*

Промпт у том случају садржи наведени контекст, циљну реч **одржати** и листу могућих значења преузетих из инвентара. Међу понуђеним кандидатима налази се, на пример, значење **ENG30-01733477-v — организовати или бити одговоран за нешто**, поред других глаголских значења као што су *продужити постојање, очувати, држати у неизмењеном стању* и слично. Од модела се затим очекује да, на основу датог контекста и понуђене листе значења, врати само идентификатор изабраног значења и кратко образложење у строго задатој структури, на пример:

```
{ "sense_id": "ENG30-01733477-v", "explanation": "U datom kontekstu glagol 'održati' odnosi se na organizovanje i sprovođenje programa, a ne na očuvanje, produžavanje trajanja ili držanje u neizmenjenom stanju." }
```

Наведени пример има искључиво илустративну функцију. Његова сврха није да документује појединачни резултат, већ да покаже како су у практичној примени биле организоване главне компоненте промпта: контекст, лема, листа кандидата значења и ограничен облик излаза.

У практичној реализацији моделу је задата улога стручњака за лексику и семантику, са задатком да, на основу контекста реченице и листе валидних значења циљне речи, идентификује значење које употребљена реч има у датом контексту. Истовремено, одговор је ограничен на строго дефинисану структуру, у којој се наводи идентификатор изабраног значења и кратко образложење избора. За потребе дисертације важнија је сама функција овог решења него његов технички облик: строго дефинисан излаз омогућава поуздано преузимање резултата и њихово укључивање у даљу полуаутоматску обраду.

Разрешење вишезначности у овом поступку спроводи се у режиму без примера, односно без навођења претходно решених примера у самом промπτу. Овај избор је детаљније теоријски образложен у одељку 2.9, те га овде није потребно поново разматрати. Овде је довољно истаћи да је такав режим омогућио једноставну и репродуктивну примену поступка над већим бројем примера.

У експериментални дизајн укључени су и генеративни језички модели и базни модели засновани на векторским репрезентацијама, како би се упоредиле две различите стратегије разрешења вишезначности: контекстуално закључивање унутар ограниченог скупа кандидата и избор на основу дистрибутивне сличности. На тај начин омогућено је да се перформансе генеративних модела сагледају у односу на јасно дефинисане референтне приступе.

Посебан значај има могућност да модел, у случајевима када ниједно од понуђених значења не одговара датом контексту, врати ознаку за ново или недостајуће значење. Овакво решење спречава да се на силу припише неко неодговарајуће постојеће значење и омогућава да се систематски издвоје примери који указују на празнине у инвентару. Такви случајеви затим могу бити предмет накнадне провере и, по потреби, основ за проширивање ресурса. У оквиру евалуације овакви случајеви су посебно бележени, јер не представљају исту врсту грешке као

избор погрешног значења у случају када одговарајуће значење постоји у инвентару. Зато су у анализи резултата издвојени као засебна категорија недодељених или непоклопљених значења.

Инвентар значења у овом истраживању има инфраструктурну улогу, јер служи као извор кандидатских значења, као посредничка структура између лексичко-семантичког ресурса и генеративног модела и као основа за накнадну проверу и редакцију анотације. Његово језгро чини ресурс **WSD-sr-sens-inv**, развијен на основу Српског ворднета, који представља основни лексичко-семантички извор и семантичку основу инвентара. Пошто Српски ворднет није обезбеђивао пуну покривеност свих лексема и значења потребних за анотацију корпуса ELEXIS-WSD-sr (одељак 2.11), инвентар је морао бити систематски проширен. За потребе експеримената коришћене су две верзије инвентара: почетна верзија, примењена у првој итерацији, и ревидирана верзија, настала након ручне провере, исправки и допуне инвентара значења.

Изградња прве верзије инвентара полазила је од енглеског ELEXIS-WSD инвентара значења, заснованог на Принстонском ворднету, и његовог поравнавања са Српским ворднетом. Након издвајања неподударних синсета, идентификован је скуп недостајућих значења релевантних за анотацију корпуса. За тај скуп су синоними и дефиниције аутоматски преведени на српски, а затим ручно кориговани. Поред тога, из корпуса су издвојене речи које носе значења (именице, глаголи, придеви и прилози), а које нису биле покривене постојећим ресурсом, па су за њих прибављени одговарајући кандидати из Принстонског ворднета, преведени и редакторски усклађени. Ради очувања хијерархијске доследности мреже додавани су и надређени синсети. За преостале јединице, нарочито термилошке изразе, новије лексеме и вишечлане изразе који нису постојали у енглеском извору, дефиниције су генерисане уз помоћ великих језичких модела, а потом ручно уређене (Krstev et al. 2025).

У евалуацију су укључене оне појаве једночланих и вишечланих лексема за које је, на основу леме и врсте речи, било могуће издвојити релевантан скуп кандидата значења. У пракси су то пре свега биле пунозначне речи: именице, глаголи, придеви, прилози и вишечлане јединице, док су функционалне вишечлане јединице разматране само ако су биле обухваћене инвентаром и анотационим слојевима.

Тако формиран инвентар не представља општи лексикографски опис српског језика, већ радни инвентар значења намењен задатку разрешења вишезначности речи у конкретном корпусу који је обрађиван (ELEXIS-WSD-sr). За сваку релевантну лему и врсту речи он обезбеђује скуп значења кандидата са идентификаторима и дефиницијама, који се затим укључује у промпт прослеђен моделу. Резултати његовог даљег проширења, кориговања, допуне и преласка са почетне на ревидирану верзију приказани су у поглављу 5.

За сваки пример речи из текста који се обрађује, из инвентара се преузима одговарајући скуп синсета, тај скуп се укључује у промпт, а затим се добијени одговор поново повезује са конкретном појавом речи у тексту. На тај начин инвентар не служи само као статички извор значења, већ и као основа за полуаутоматску изградњу и редакцију корпуса анотираног значењима. Схематски приказ радног тока разрешења вишезначности речи у предложеном полуаутоматском поступку дат је на Слици 3.3.



Слика 3.3 Схематски приказ радног тока разрешења вишезначности речи применом генеративног модела

Поступак се сажето може представити као обављање неколико повезаних корака. Из корпуса се издваја конкретна појава речи или вишечлане јединице. На основу леме и врсте речи, из инвентара се преузима скуп кандидатских значења. Тај скуп се, заједно са реченицом у којој је циљна јединица експлицитно означена, уноси у структурирани промпт. Модел затим врши избор најприкладнијег значења, односно семантички најближег кандидата, и враћа одговор у строго задатој структури. Потом се резултат повезује са конкретном појавом у корпусу и укључује у даљи ток анотације, ручне провере и, по потреби, корекције и допуне инвентара

значења. Поступак се не спроводи једнократно, већ има итеративни карактер, у оквиру кога се резултати аутоматске анотације накнадно ручно проверавају, а затим користе за даље проширивање и унапређење инвентара значења (Слика 3.4).



Слика 3.4 Итеративни процес аутоматске анотације, ручне корекције и допуне инвентара значења

У оквиру предложене методологије потребно је разликовати почетно издвајање кандидата на нивоу корпуса од ручне провере контролних подскупова. У првој фази, након примене почетне верзије инвентара на корпус ELEXIS-WSD-sr, издвојене су све појаве за које је модел вратио ознаку **NEW_SENSE** јер те појаве вероватно указују да у постојећем инвентару можда недостаје одговарајуће значење. Ова фаза представљала је једнократно, аутоматски подржано претраживање корпуса и послужила је за ручну допуну инвентара WSD-sr-sens-inv-v2 новим значењима. Она, међутим, не подразумева завршену ручну обраду целокупног корпуса, већ почетни корак у методологији поступног развоја ресурса у условима ограничене доступности претходно аотираних података.

Другу компоненту поступка чинила је ручна провера контролних подскупова реченица. Првих 120 реченица имало је улогу развојно-аналитичког подскупа: на њима су проверени излази модела, уочени случајеви погрешно додељених или непокривених значења и издвојени кандидати и запажања за даље унапређење инвентара. После допуне ресурса, исти подскуп омогућио је поређење резултата између две итерације, док је накнадна провера на 300 реченица коришћена као додатна контролна тачка. Нови случајеви уочени у том проширеном скупу бележени су као основ за наредне допуне, али нису коришћени за накнадно прилагођавање инвентара у истом евалуационом кругу. Током даљег истраживања ће се радити и додатни круг евалуације. На тај начин поступак задржава полуаутоматски карактер: аутоматска анотација убрзава откривање празнина у инвентару, док ручна провера обезбеђује квалитет, упоредивост и основу за наредне итерације.

У оквиру овог поступка генеративни модел не делује као самостални доносилац коначних лексикографских одлука, већ као средство за убрзавање и структурирање полуаутоматске анотације. Његова улога је да, на основу контекста и унапред припремљеног скупа значења,

понуди највероватнији избор и кратко образложење. Вредност таквог излаза не лежи само у појединачном аутоматском избору, већ и у томе што омогућава брже обогаћивање репозиторијума, уочавање недостатака у ресурсу и организовање експертске провере над већ припремљеним предлозима. Зато разрешење вишезначности речи применом генеративног модела у овој дисертацији представља централни део методолошког поступка за повезивање корпуса, лексичко-семантичког ресурса и полуаутоматске анотације.

3.4 Разрешење вишезначности речи – дестилација

Поред поступка описаног у претходном одељку, у овом истраживању разматра се и могућност преноса добијеног знања на мањи и рачунски ефикаснији модел. У овом контексту, термин дестилација модела користи се у ширем практичном смислу, као пренос знања из сложенијег поступка анотације на мањи модел посредством аутоматски означеног скупа података, а не нужно у ужем смислу класичне дестилације засноване на неким излазима наставничког модела. Полазна идеја је да се скуп података добијен полуаутоматском анотацијом применом великог језичког модела искористи као основа за обуку мањег модела, погоднијег за бржу и јефтинију примену на већем броју примера. У том смислу, под дестилацијом модела овде се подразумева поступак у коме се одлуке и структуре добијене сложенијим поступком анотације преносе на мањи модел посредством аутоматски означеног скупа података.

За разлику од поступка из одељка 3.3, чији је један од исхода и допуна инвентара значења, мањи модел нема улогу у развоју самог ресурса. Његова сврха је другачија: да на основу већ припремљеног инвентара и претходно анотираног скупа, омогући брже разрешење вишезначности речи у новим примерима. На тај начин уводи се јасно раздвајање између два функционално различита нивоа система: већи модел користи се за полуаутоматску анотацију и унапређење ресурса, док се мањи модел обучава за ефикаснију каснију примену над већим бројем реченица.

За обуку мањег модела користи се сребрни скуп (енгл. *silver dataset*), односно скуп реченица аутоматски анотираних у поступку описаном у претходном одељку. Златни скуп (енгл. *gold dataset*) ручно проверених реченица не користи се за обуку, већ искључиво за завршну евалуацију. Овакво раздвајање омогућава да се процена квалитета мањег модела заснива на независном ручно верификованом скупу, а да се обука истовремено ослони на већи број аутоматски добијених примера.

Формулација задатка за мањи модел не заснива се на једној глобалној вишекласној класификацији над свим значењима у инвентару, већ на упаривању реченице и појединачног значења из инвентара значења. Избор значења ни у овом поступку не врши се над целокупним инвентаром значења, већ је ограничен на скуп кандидата значења која одговарају датој лем и врсти речи. За сваки пример најпре се, као и у претходном поступку, из реченице издваја циљна лексема, одређују се њена лема и врста речи и из инвентара се преузима скуп одговарајућих кандидата значења. Затим се за сваки кандидат значења формира пар који чине:

- 1) реченица са експлицитно означеном циљном јединицом
- 2) дефиниција потенцијално одговарајућег значења.

На тај начин задатак се трансформише у процену семантичке сагласности између конкретне употребе циљне речи у реченици и дефиниције једног могућег значења. Позитивни парови су они у којима дефиниција одговара значењу које је већ додељено циљној јединици у сребрном скупу, док су негативни парови они у којима је кандидатско значење везано за исту лему, али

не одговара датом контексту. Таква организација скупа за учење омогућава да се модел обучава као модел за рангирање или семантичко упаривање, а не као затворени класификатор фиксног скупа ознака.

Оваква поставка је нарочито погодна за modele типа *sentence transformer*, јер омогућава да се реченични контекст и дефиниција значења представе у заједничком векторском простору. У току обуке модел учи да репрезентације позитивних парова буду међусобно ближе, а репрезентације негативних парова међусобно удаљеније у заједничком векторском простору. У фази примене, за дату реченицу и циљну јединицу из инвентара значења се преузимају сви кандидати који одговарају леми и врсти речи, затим се формира по један пар од реченице и сваког од тих значења, а модел за сваки пар израчунава степен сличности. Коначно, могућа значења се рангирају према добијеном резултату, а као избор се узима највише рангирано значење. На овај начин добија се не само коначна одлука, већ и рангирана листа кандидата, што је методолошки корисно јер омогућава увид у блиска и конкурентна значења.

Улаз за мањи модел, дакле, задржава кључне компоненте коришћене и у поступку са великим језичким моделом: реченични контекст, означену циљну јединицу и дефинисана значења у инвентару. Међутим, уместо да се задатак реализује у облику промпта прослеђеног генеративном моделу, овде се он трансформише у стандардизован проблем семантичког упаривања. То омогућава да се модел обучи једном, а затим примењује локално, брже и уз знатно мање рачунске трошкове.

Предност оваквог приступа лежи пре свега у практичности. Мањи модел је бржи, захтева мање рачунских ресурса и самим тим погоднији је за обраду већих количина нових података него поступак који се ослања на већи генеративни модел. Треба напоменути да је обрада једне реченице великим језичким моделом трајала у просеку око три минута, што такав приступ чини неприменљивим за велике корпусе. Уз то, не треба занемарити ни финансијске импликације (трошкове), који за велике корпусе не би били прихватљиви. Истовремено, задржава се веза са инвентаром значења, јер се одлука и даље доноси унутар скупа значења кандидата дефинисаних за дату лему и врсту речи. На тај начин систем остаје контролисан и интерпретабилан, а не своди се на слободно предвиђање без ослонаца у лексичко-семантичком ресурсу.

Ограничење овог поступка, међутим, јесте у томе што мањи модел не доприноси непосредно даљој изградњи инвентара значења. За разлику од већег генеративног модела, он не служи за уочавање нових или недостајућих значења, нити за формулисање нових дефиниција. Његова улога је ужа: да на основу већ постојећег и претходно ревидираног ресурса што ефикасније процени које од понуђених значења најбоље одговара датом контексту. Управо зато овај поступак не треба посматрати као замену за приступ описан у одељку 3.3, већ као његову оперативну допуну, усмерену ка бржој и економичнијој примени већ изграђеног ресурса.

Коначна евалуација мањег модела врши се на ручно провереном златном скупу, који није коришћен у обуци и зато представља независну основу за процену квалитета модела. На тај начин може се непосредно проценити у којој мери је знање добијено полуаутоматском анотацијом и организовано кроз инвентар значења успешно пренето на бржи и мањи модел. У том поређењу не очекује се нужно да мањи модел достигне квалитет већег генеративног модела у свим случајевима, али се испитује да ли добијени компромис између тачности, брзине и рачунске цене чини такав модел употребљивим у практичним сценаријима примене.

3.5 Превод синсета

У оквиру овог истраживања аутоматски превод синсета посматра се као засебан методолошки проблем чији је циљ да се испита у којој мери велики језички модели могу да подрже практично проширивање Српског ворднета превођењем његових кључних лексичко-семантичких компоненти са енглеског на српски језик. Полазиште овог приступа чини чињеница да су релације међу синсетима, њихови идентификатори и укупна структура мреже у основи језички независни, док се језички специфичан део синсета пре свега односи на глосу и скуп литерала. Због тога је у овом истраживању тежиште постављено управо на превођење дефиниција и литерала, док се остали структурни елементи синсета третирају као стабилна формална основа која се у поступку превођења непосредно пресликава. У српском ворднету нису уношени примери употребе свуда систематично, тако да је та допуна једна од планираних будућих активности.

Основни циљ није био само да се добију српски преводни еквиваленти, већ и да се испита у којој мери различито организовани радни токови могу да обезбеде резултате који су лексички прихватљиви, семантички прецизни и употребљиви за даљу лексикографску обраду. У том смислу, превод синсета није посматран као изолован задатак машинског превођења, већ као контролисан поступак трансформације лексичко-семантичког записа из једног језика у други, при чему је неопходно очувати значење, унутрашњу кохерентност синсета и његову употребљивост у оквиру циљног ресурса.

Методолошки посматрано, у истраживању су разматрана три радна тока различите сложености. Сва три полазе од истог објекта обраде, односно синсета као целине, али се разликују по томе који се његови елементи у појединим фазама експлицитно укључују у обраду, као и по степену контроле над процесом формирања коначног резултата. Основна логика оваквог поређења јесте да се утврди да ли је за практичан рад довољан једноставан модел директног превођења, или је за добијање квалитетнијих резултата неопходно увести сложеније, вишефазне поступке са експлицитнијом семантичком контролом.

Први радни ток представља базни, односно полазни приступ. У њему се моделу задају само енглеска глоса и скуп енглеских литерала, а од њега се захтева да врати превод глосе и одговарајући скуп српских литерала. Овај приступ има јасну методолошку улогу, јер служи као референтна тачка за поређење са сложенијим решењима. Његова предност лежи у организационој једноставности и могућности брзог добијања резултата. Истовремено, управо због те једноставности, он је изложен повећаном ризику од губитка семантичке прецизности, нарочито у случајевима у којима кратка глоса није довољна да поуздано усмери модел ка жељеном значењу, или када директно превођење литерала не обезбеђује њихову пуну усклађеност са значењем синсета као целине.

Други радни ток уводи вишефазну обраду, у којој се превод не добија једним кораком, већ се поступак разлаже на више међусобно повезаних фаза. У почетном кораку модел анализира изворни синсет на енглеском језику, са циљем да издвоји његове кључне семантичке одлике, домен (који се преузима из енглеског синсета) и разликоване карактеристике значења. Овај корак има функцију интерне семантичке стабилизације, односно обезбеђивања да се даљи превод не ослања искључиво на површински текст глосе, већ и на експлицитније разумевање самог концепта.

Након тога следи превођење глосе на српски језик, али не као изолованог низа речи, већ као дефиниције која треба да остане верна претходно анализираном значењу. Потом се засебно преводе и литерали, при чему се не полази од претпоставке да је сваки директан превод уједно

и добар кандидат за коначни синсет. У наредном кораку полазни скуп српских кандидата проширује се новим литеалима који би такође могли одговарати датом значењу, након чега се врши њихово филтрирање. На тај начин овај радни ток не третира превод синсета као питање једноставне међујезичке еквиваленције, већ као поступак у коме се најпре формира шири скуп потенцијалних решења, а затим се из њега издвајају само она која су довољно прецизна и употребљива у оквиру циљног ресурса.

Важан елемент овог другог радног тока представља и завршно преформулисање глосе, како би она била што прикладнија и у духу записа у ворднету. То подразумева да глоса буде јасна, сажета и семантички прецизна, те да нема кружних дефиниција, односно да не дефинише значење ослањањем на исте или готово исте лексеме које се јављају у скупу литеала. На тај начин се у поступак не уводи само контрола превода, већ и контрола квалитета саме лексикографске формулације.

Кључна разлика између другог и трећег радног тока није само у броју корака, већ у природи посредне семантичке репрезентације. У другом току анализа синсета служи као контекст за низ одвојених операција — превод глосе, превод и проширење литеала, њихово филтрирање и завршно преформулисање глосе — док се у трећем току тежиште помера ка експлицитно формулисаној проширеној дефиницији, која преузима улогу централног посредног корака у даљем обликовању српског записа.

Трећи радни ток представља методолошки најсложенији приступ. Он полази од претпоставке да кратка глоса, чак и када је правилно преведена, не мора увек бити довољна основа за добијање најбољег могућег српског записа. Због тога се у овом приступу најпре формира шира, експлицитнија и садржински богатија репрезентација синсета. Та репрезентација не заснива се само на оригиналној глоси и литеалима, већ и на другим расположивим елементима синсета, као што су примери употребе, доменске информације и семантичке релације. На основу таквог скупа података модел најпре формулише проширену дефиницију на енглеском језику, која има улогу посредне семантичке репрезентације.

Сврха ове проширене дефиниције није да непосредно постане глоса у циљном ресурсу (на пример српском), већ да омогући експлицитнију артикулацију значења пре него што се крене на формулисање српског пандана. Након тога се ова проширена дефиниција преноси на српски језик, а затим се на основу ње генеришу кандидатски српски литеали. У овој фази посебно је важно да модел не нуди само формално могуће преводе, већ и да разликује решења која су заиста лексикализована и природна у српском језику од оних која имају карактер описних или привремених парафраза. Из тог разлога се у даљем кораку врши селекција литеала, при чему се задржавају само они кандидати чија су значења истовремено довољно прецизна и језички природна. На тај начин се део посла који је у другом радном току реализован кроз засебне фазе превода, проширивања и редукције литеала у већој мери преноси на ниво саме посредне дефиниције, из које се кандидати за литеале затим непосредније изводе.

Тек након избора коначног скупа литеала приступа се формулисању завршне српске глосе. На тај начин је у трећем радном току редослед обрнут у односу на базни модел: глоса више није прва и непосредна последица енглеског улаза, већ коначни резултат поступка у коме су претходно стабилизовани значење, обим концепта и лексички избор. Овакво решење ближе је традиционалној лексикографској пракси, јер полази од концепта и његове анализе, а тек потом приступа коначном обликовању дефиниције и скупа литеала.

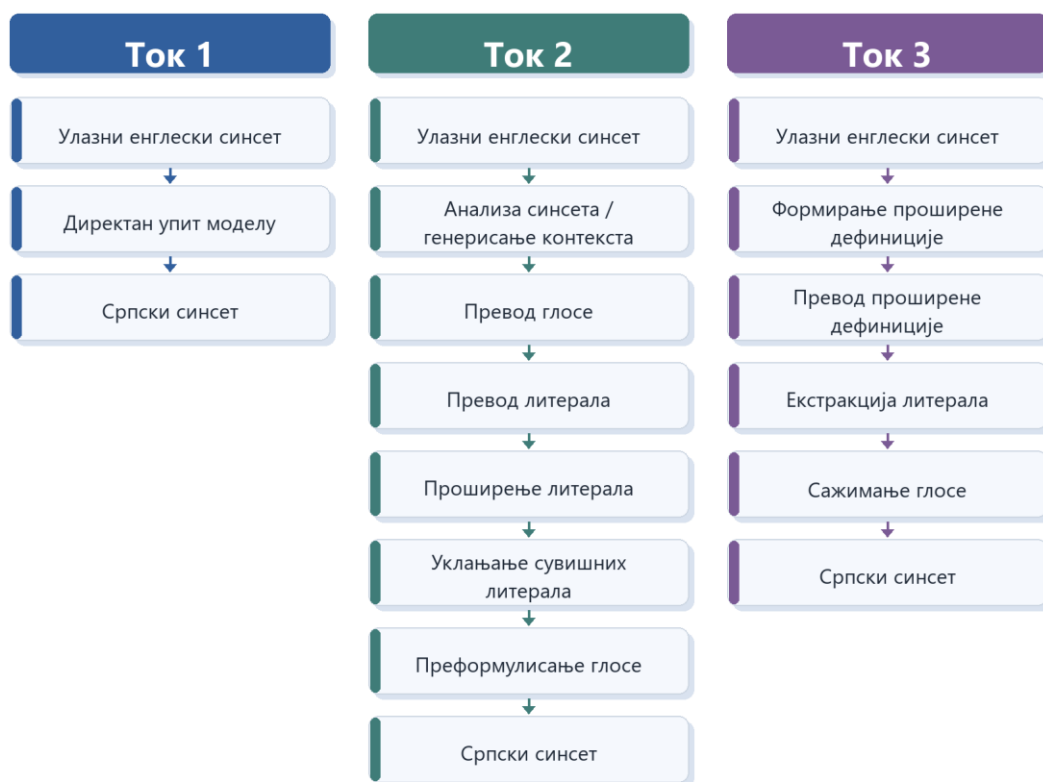
Завршни корак овог приступа обухвата интерну валидацију резултата. У оквиру те провере анализира се усклађеност између изабраних литеала, завршне глосе и изворне семантичке

структуре синсета. Посебна пажња посвећује се питањима кружних дефиниција, семантичког померања, прешироких или преуских еквивалената, као и уопште погодности добијеног резултата за укључивање у ворднет. Иако се ова провера и даље ослања на аутоматизован поступак, њена функција је методолошки блиска лексикографској редакцији, јер не служи само да потврди да је нешто преведено, већ и да процени да ли је добијени синсет унутрашње доследан и употребљив (према претходно дефинисаним правилима).

Поред интерне валидације у оквиру самог радног тока, за процену квалитета добијених српских синсета предвиђена је и ручна експертска провера. Њена улога није да понови аутоматизовану проверу, већ да процени лексикографску и језичку употребљивост добијеног резултата у српском ворднету. Проверу врши један експерт, на основу структурираног обрасца у коме се одвојено процењују глоса и литерали. За глосу се проверава да ли је формулисана у духу српског језика, да ли адекватно одражава значење синсета, да ли омогућава разликовање од сродних значења, да ли је њена структура сагласна врсти речи и да ли је формулисана у складу са основним лексикографским принципима.

У процени литерала посебно се бележи да ли сви предложени литерали припадају датом значењу, да ли међу њима има сувишних јединица и да ли недостају лексеме које би требало укључити у синсет. Ако је потребно, експерт уноси предлог кориговане глосе, као и конкретне литерале који недостају или су вишак, тј. не припадају синсету. На тај начин ручна провера не служи за доделу нумеричке оцене квалитета, већ и као механизам за прикупљање редакторских интервенција које се могу употребити у наредној фази уређивања и допуне ресурса.

Добијене експертске оцене имају и додатну методолошку улогу: оне служе као основ за избор радног тока који је најпогоднији за даљу примену у развоју Српског ворднета. Пошто паралелно одржавање три различита поступка није практично у продукционој фази, ручна процена омогућава да се утврди који приступ даје најбољи однос између квалитета, сложености и употребљивости резултата. Уколико ниједан од испитаних радних токова не покаже довољну предност у свим аспектима, резултати евалуације могу послужити као основ за формулисање новог, комбинованог или унапређеног поступка, у коме би се задржале најуспешније компоненте појединачних приступа. Ради јаснијег увида у разлике између описаних приступа, на Слици 3.5 дат је упоредни приказ три радна тока, који се разликују по степену експлицитности семантичке обраде, броју корака и начину формирања српског синсета.



Слика 3.5 Приказ радних токова за допуну Српског ворднета

Упоредно испитивање ова три радна тока омогућава да се превод синсета не сагледа само као питање избора модела, већ и као питање организације самог поступка. На једној страни налази се директно превођење као најједноставнији и најбржи модел обраде, док се на другој страни налазе сложенији радни токови који уводе додатне кораке семантичке анализе, генерисања кандидата, селекције и валидације. Основна претпоставка јесте да повећање унутрашње структурираности поступка може довести до квалитетнијих и лексикографски прихватљивијих резултата, али по цену веће организационе и рачунске сложености.

Истраживачки задатак зато није био само да се покаже да велики језички модели могу да произведу српске преводе глоса и литерала, већ и да се утврди који тип радног тока обезбеђује најповољнији однос између квалитета резултата, сложености поступка и практичне употребљивости у даљем раду на Српском ворднету. У овом истраживању превод синсета стога није третиран као помоћна техничка активност, већ као засебан експериментални и методолошки оквир, релевантан за полуаутоматско проширивање лексичко-семантичких ресурса за српски језик.

3.6 Поравнање лексичко-семантичких ресурса уз употребу генеративних модела

У овом истраживању примена генеративних језичких модела не усмерава се на потпуно аутоматско поравнање целокупних ресурса, већ на контролисано испитивање могућности оваквог приступа на ограниченом подскупу података. Из тог разлога се анализа спроводи над издвојеним делом Речника савременог српског језика (Stanković et al. 2025) и одговарајућим

делом српског ворднета, чиме се омогућава детаљнија квалитативна и квантитативна анализа резултата.

Полазни корак подразумева генерисање свих релевантних парова између речничких дефиниција и синсета. За сваку лексему издвајају се све речничке ставке, као и сви синсети у којима се та лексема јавља као литерал. На тај начин се формира скуп кандидата који представља декартов производ речничких значења и одговарајућих синсета, чиме се обезбеђује да ниједно потенцијално поравнање не буде унапред искључено.

Да би се обезбедила информативнија и репрезентативнија основа за поређење, глосе синсета се проширују. Проширена дефиниција обухвата не само основну глосу, већ и оне елементе семантичке мреже који могу допринети прецизнијем разграничењу значења, пре свега доменске ознаке и релевантне семантичке релације, као што су хиперонимија и меронимија. На тај начин се ублажава проблем сажетости глоса и моделу се пружа богатији контекст за доношење одлуке.

Задатак се затим формулише као класификација односа између речничке дефиниције и синсета. За сваки пар модел одређује једну од следећих категорија: потпуно поклапање, уже значење, шире значење, сродност или одсуство релевантне везе. Поред саме класификације, од модела се може захтевати и кратко образложење, које служи као додатни увид у процес одлучивања, али се не користи као примарни евалуациони сигнал.

Посебна пажња посвећује се анализи добијених резултата. Резултати се не тумаче искључиво као излаз система за поравнање, већ и као индикатор структурних и лексикографских ограничења ресурса. На пример, случајеви у којима једна речничка дефиниција одговара већем броју синсета могу указивати на различиту грануларност значења, док ситуације у којима ниједан синсет не одговара дефиницији из речника могу указивати на непотпуну покривеност ворднета. Слично томе, више дефиниција из речника које се поклапају са једним синсетом могу указивати на преклапање или недовољну диференцијацију значења у речнику.

На тај начин генеративни модел се не користи искључиво као средство за аутоматско поравнање, већ и као аналитички инструмент за идентификацију недостатака и потенцијалних унапређења лексичко-семантичких ресурса. Овај приступ омогућава итеративно побољшање и речника и ворднета, чиме се превазилази ограничење посматрања поравнања као једнократног техничког поступка.

3.7 Екстерна евалуација обогаћеног Српског ворднета у анализи сентимента

Да би се испитало да ли вредности сентимента додељене синсетима Српског ворднета имају практичну вредност у задатку аутоматске анализе сентимента, у раду се предвиђа екстерна евалуација над реченицама са познатом ознаком сентимента које ће бити представљене у одељку 4.7. Полазни скуп чине реченице аотиране према поларитету сентимента, док се у овом истраживању над истим реченицама додаје лексичко-семантички слој аотације. За сваку лексему која носи значење у реченици одређују се лема и врста речи, затим се из инвентара значења издвајају кандидати синсета, а потом се поступком за разрешење вишезначности речи бира синсет који најбоље одговара датом контексту.

Оваква организација података омогућава директно поређење два начина рачунања сентимента реченице. У првом, лексичком приступу, сентимент се рачуна на основу поларитета лема присутних у реченици. У другом приступу заснованом на значењима, сентимент се рачуна на

основу поларитета синсета који су додељени конкретним примерцима речи у контексту. На тај начин исти текстуални пример може бити обрађен и без информације о значењу и са експлицитном информацијом о значењу, што омогућава изоловану процену доприноса структуре ворднета и разрешења вишезначности речи.

За потребе поређења дефинишу се три основне конфигурације. Прва конфигурација представља полазну поставку која се ослања на вредности сентимента лема из ресурса Senti-pol-sr. У овом ресурсу поларитет леме представљен је дискретном вредношћу из скупа $\{-1, 0, 1\}$, где -1 означава негативну, 0 неутралну, а 1 позитивну оријентацију сентимента. Друга конфигурација најпре разрешењем вишезначности свакој лемидодељује одговарајући синсет, а затим задржава полазне вредности сентимента тог синсета, добијене пресликавањем из енглеског SentiWordNet-a. Трећа конфигурација користи синсете који одговарају лемидобијене разрешењем вишезначности и коначну обогачену верзију Српског ворднета, односно проширени ресурс са вредностима сентимента добијеним методологијом развијеном у овој дисертацији. Конфигурација сваког синсета има две компоненте сентимента, POS и NEG, чије су вредности у интервалу $[0, 1]$ и чији је збир ≤ 1 .

Сентимент реченице рачуна се транспарентним поступком агрегирања. За сваку реченицу сабирају се доприноси релевантних лексичких јединица, при чему се у лексичком приступу користи дискретни поларитет леме, а у приступу заснованом на значењима разлика позитивне и негативне вредности изабраног синсета. Пошто лексичка полазна поставка и синсетске конфигурације користе различите типове вредности сентимента, резултати се не пореде на нивоу сирових скорова, већ на нивоу коначних класа реченица. Да би се смањио утицај дужине реченице и различите покривености ресурса, у обе поставке користи се просечна вредност преко оних јединица за које је пронађена одговарајућа информација везана за сентимент. Једноставност ове процедуре представља намерну методолошку одлуку: циљ није изградња општег система за анализу сентимента српског језика, већ испитивање да ли прецизније лексичко-семантичко представљање речи доводи до мерљивог побољшања у односу на приступ заснован само на лемама.

За сваку реченицу r посматра се скуп релевантних садржајних јединица $T(r)$, који обухвата лексеме укључене у рачунање сентимента. У лексичком приступу користи се подскуп $T_{\text{lemma}}(r)$, који садржи јединице за које је пронађена вредност сентимента у лексикону Senti-pol-sr. Просечни лексички скор реченице дефинише се као:

$$\text{score}_{\text{lemma}}(r) = \frac{1}{|T_{\text{lemma}}(r)|} \sum_{t \in T_{\text{lemma}}(r)} \text{pol}(\text{lemma}(t)).$$

У овој формули $\text{pol}(\text{lemma}(t))$ означава дискретну лексичку вредност сентимента за одговарајућу лему, при чему важи:

$$\text{pol}(\text{lemma}(t)) \in \{-1, 0, 1\}.$$

Ако у реченици нема јединица покривених лексиконом, реченица се у овој конфигурацији може означити као неутрална или искључити из анализе, у зависности од унапред дефинисаног евалуационог протокола.

У приступу заснованом на значењима користи се подскуп $T_{\text{synset}}(r)$, који садржи јединице којима је додељен синсет са доступним вредностима сентимента. Свакој

јединици t придружује се синсет s_t , изабран поступком разрешења вишезначности речи. Просечни скор сентимента реченице базиран на придруженим синсетима дефинише се као:

$$score_{\text{synset}}(r) = \frac{1}{|T_{\text{synset}}(r)|} \sum_{t \in T_{\text{synset}}(r)} (POS(s_t) - NEG(s_t)).$$

У овој формули $POS(s_t)$ и $NEG(s_t)$ означавају позитивну и негативну компоненту сентимента синсета s_t , при чему важи:

$$POS(s_t), NEG(s_t) \in [0,1].$$

Ако се позитивна и негативна компонента чувају одвојено, на нивоу реченице могу се дефинисати просечне компоненте:

$$POS(r) = \frac{1}{|T_{\text{synset}}(r)|} \sum_{t \in T_{\text{synset}}(r)} POS(s_t),$$

$$NEG(r) = \frac{1}{|T_{\text{synset}}(r)|} \sum_{t \in T_{\text{synset}}(r)} NEG(s_t).$$

Коначна ознака сентимента реченице одређује се на основу добијеног скорa и унапред дефинисаног прага θ . У лексичком приступу, због дискретне природе улазних вредности, може се користити следеће правило:

$$label_{\text{lemma}}(r) = \begin{cases} \text{positive}, & score_{\text{lemma}}(r) > 0, \\ \text{negative}, & score_{\text{lemma}}(r) < 0, \\ \text{neutral}, & score_{\text{lemma}}(r) = 0. \end{cases}$$

У приступу заснованом на синсетима користи се праг који омогућава да се слабе разлике између позитивне и негативне компоненте третирају као неутралне:

$$label_{\text{synset}}(r) = \begin{cases} \text{positive}, & score_{\text{synset}}(r) > \theta, \\ \text{negative}, & score_{\text{synset}}(r) < -\theta, \\ \text{neutral}, & -\theta \leq score_{\text{synset}}(r) \leq \theta. \end{cases}$$

Праг θ не подешава се на тест скупу, већ се бира на развојном скупу на основу најбоље вредности мере $macro\text{-}F1$. У ту сврху пореде се унапред дефинисане кандидатске вредности, са гушћим распоредом око нуле, на пример 0, 0,01, 0,02, 0,04, 0,06, 0,10, 0,20, 0,33 и 0,50. Изабрани праг се затим примењује на тест скуп без додатног прилагођавања, чиме се избегава накнадно оптимизовање резултата према тест скупу и обезбеђује фер поређење различитих конфигурација.

Поред конфигурације засноване на синсету изабраном поступком разрешења вишезначности речи, уведена је и контролна конфигурација која не користи информацију о конкретно изабраном значењу. У тој поставци, означеној као avg , за сваку лему и врсту речи рачуна се просек вредности $POS(s)$ - $NEG(s)$ преко свих синсета у ворднету који садрже ту лему и исту врсту речи. Ова контрола омогућава да се раздвоје два фактора: квалитет вредности сентимента у самом ресурсу и ограничења тренутног инвентара значења и поступка

разрешења вишезначности речи. Ако avg конфигурација постиже бољи или упоредив резултат од WSD конфигурације, то не значи да је разрешење вишезначности непотребно, већ указује да тренутна покривеност и повезаност инвентара значења са слојем који укључује сентименте још увек ограничавају добитак од избора конкретног синсета.

Основне конфигурације за поређење у задатку анализе сентимента приказане су у Табели 3.1.

Табела 3.1 Конфигурације за екстерну евалуацију у анализи сентимента

Ознака	Опис	Шта се проверава
L0	Полазни лексикон је заснован на `Senti-pol-sr`; користе се леме, без WSD-a.	Полазна вредност лексичког приступа.
S0	Полазна поставка заснована на синсетима; користе се синсети добијени разрешавањем вишезначности и вредности сентимента пресликане из енглеског SentiWordNet-a.	Ефекат преласка са леме на значење уз постојећи слој сентимената.
S1-7-wsd	Коначна обogaћена верзија Српског ворднета; користе се синсети добијени разрешавањем вишезначности и вредности сентимента добијене методологијом развијеном у дисертацији.	Пун ефекат обogaћеног ресурса у односу на полазно пресликавање из SentiWordNet-a.
S1-7-avg	Опциона контрола без разрешавања вишезначности; за вишезначне леме користи се просек преко кандидатских синсета.	Да ли је разрешавање вишезначности неопходно или је довољно само постојање лексикона синсета, тј. ворднета.

3.8 Циљана провера употребе слоја за разрешење вишезначности речи у машинском превођењу

Поред евалуације у задатку анализе сентимента, у раду је уведена и циљана провера употребљивости слоја за разрешење вишезначности речи у машинском превођењу. Овај део експеримента не третира се као изградња или евалуација општег система за машинско превођење, већ као контролисана провера да ли експлицитно навођење значења може помоћи при превођењу лексички двосмислених јединица. Због тога се евалуација не заснива на општим мерама квалитета превода, већ на провери избора преводног еквивалента за унапред означене циљне лексичке јединице.

Као улаз се користе реченице из српског дела ELEXIS-WSD-a за које постоји анотација значења речи повезана са инвентаром значења. Смер превођења је са српског на енглески језик. Овакав смер је методолошки оправдан зато што се разрешење вишезначности врши на српском тексту, док енглески део паралелног корпуса и енглески литерали одговарајућег синсета омогућавају проверу да ли је модел изабрао одговарајући преводни еквивалент.

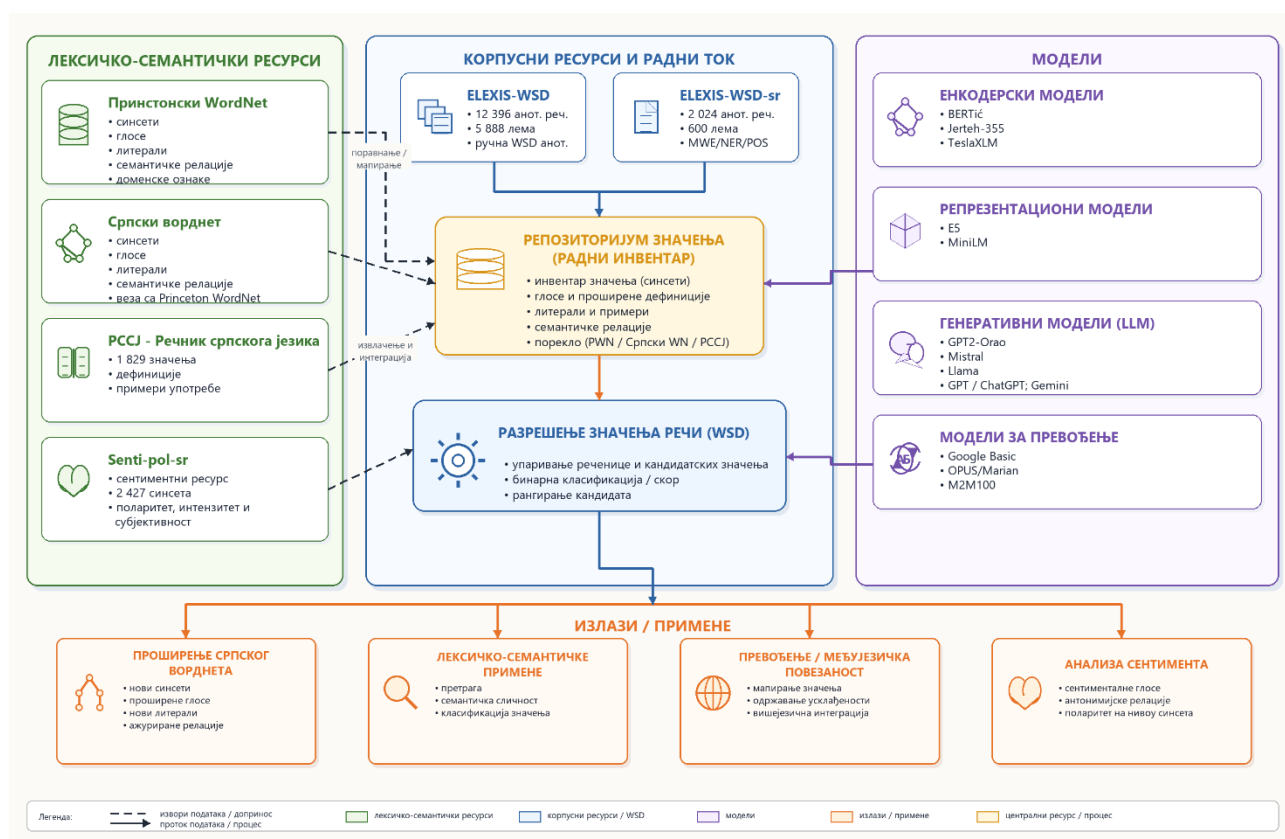
Примери се бирају из реченица које садрже бар једну означену лексему са више значења кандидата у инвентару значења. За ову проверу посебно су погодни примери у којима је изабрано значење преко ILI-ја или одговарајућег синсета у Принстонском ворднету повезано са енглеским литеаралима који се могу употребити као преводни траг.

За сваку изабрану реченицу формирају се две поставке улаза. У првој поставци модел добија само српску реченицу и задатак да је преведе на енглески. У другој поставци модел, поред исте реченице, добија кратку напомену о значењу у облику преводних трагова: уз сваку означену лексему наводи се списак једног или више енглеских литерала из енглеског синсета који је преко ILL-ја повезан са изабраним српским синсетом. У напомену не треба уносити идентификатор синсета, глосу или шири лексикографски опис, јер је циљ да се моделу понуди минимална информација корисна за избор преводних еквивалената. На тај начин се испитује разлика између превођења без експлицитне семантичке информације и превођења у коме су моделу понуђени кратки трагови из међујезички повезаног ресурса који се користи (ворднета).

Евалуација се спроводи по реченицама, уз проверу превода означених лексичких јединица. За сваку реченицу бележи се да ли су означене лексеме преведене у складу са означеним значењима. Као основа за процену користе се ELEXIS-WSD-sr анотације, енглески део паралелног корпуса и енглески литерали одговарајућих синсета. Исходи се разврставају у четири категорије: случајеви у којима напомена о значењу исправља погрешан превод, случајеви у којима оба превода остају тачна, случајеви у којима напомена о значењу не мења погрешан превод и случајеви у којима напомена о значењу погоршава превод.

4 Коришћени скупови података и велики језички модели

У овом поглављу биће представљени коришћени скупови података и велики језички модели који чине експериментално окружење овог истраживања. Скупови података обухватају лексичко-семантичке ресурсе и корпусе који обезбеђују структурирано знање о значењима и њиховој употреби, док модели обухватају како базе, претходно обучене моделе, тако и њихове фино подешене варијанте прилагођене српском језику. На тај начин успоставља се јасна веза између методолошког оквира изложеног у претходном поглављу и резултата приказаних у поглављима 5 и 6.



Слика 4.1 Преглед коришћених ресурса, скупова података и модела у експерименталном окружењу

Слика 4.1 приказује међусобну повезаност лексичко-семантичких ресурса, корпусних скупова података и модела примењених у овом истраживању. Лексички ресурси, као што су Принстонски WordNet, Српски ворднет, речник српског језика и ресурс за обележавање сентимента Senti-pol-sr, обезбеђују основу за формирање радног инвентара значења. Корпусни ресурси ELEXIS-WSD и ELEXIS-WSD-sr служе као извор ановираних примера и података за задатак разрешења вишезначности речи, док се различите групе модела користе за представљање, поређење и избор значења. На тај начин приказује се како се полазни ресурси повезују са поступком разрешења вишезначности речи и са каснијим применама у проширењу Српског ворднета, лексичко-семантичким применама, међујезичком повезивању и анализи сентимента.

4.1 Принстонски ворднет

Принстонски ворднет у овом раду има улогу полазног лексичко-семантичког ресурса за дефинисање и организовање скупа значења. Иако су његова структура и теоријске основе већ размотрене у претходном поглављу, на овом месту он се посматра из перспективе примене у експерименталном делу истраживања. У том смислу, користи се као извор синсета, глоса, примера употребе и семантичких релација који омогућавају формализацију значења и њихову даљу обраду (Miller 1995; Fellbaum 1998).

У практичном смислу, из Принстонског ворднета преузимају се глосе, литерали, примери употребе и одабране релације између синсета. Посебно су релевантне хиперонимијске и хипонимијске релације, које омогућавају позиционирање значења у хијерархијској структури, као и меронимијске и холонимске релације, које описују односе део–целина. У појединим случајевима користе се и доменске ознаке, када су доступне, ради прецизнијег одређивања семантичког поља значења.

Значај овог ресурса у области обраде природног језика вишеструко је потврђен. Принстонски ворднет је током деценија постао стандардни ресурс за представљање лексичко-семантичког знања о енглеском језику, а његова структура послужила је као основа за развој бројних вишејезичких и националних ворднета, као и за задатке као што су разрешење вишезначности речи, моделирање семантичких односа и изградња лексичко-семантичких репрезентација (Miller 1995; Fellbaum 1998). Његова релевантност задржана је и у новијим истраживањима у којима се лексички ресурси доводе у везу са савременим језичким моделима и ВЈМ (Almeman et al. 2024).

У оквиру овог истраживања, Принстонски ворднет обезбеђује почетни инвентар значења који се доводи у везу са одговарајућим јединицама у Српском ворднету. Глосе и примери употребе из овог ресурса користе се као полазни материјал за генерисање проширених дефиниција, превод, адаптацију синсета и даљу проверу у српском језичком окружењу. Семантичке релације имају додатну улогу у проширивању контекста значења и у прецизнијем позиционирању појединачних лексема унутар шире концептуалне структуре.

Посебан значај Принстонског ворднета у овом раду произлази из његове улоге у проширивању Српског ворднета. Како је Српски ворднет развијан у оквиру ширег пројекта и повезан са Принстонским ворднетом, овај ресурс омогућава систематично уочавање недостајућих значења, идентификацију потенцијалних допуна и очување структурне усклађености приликом проширивања (Krstev et al. 2004; Stanković, Mladenović, et al. 2018).

Истовремено, употреба Принстонског ворднета у контексту српског језика подразумева и одређена ограничења. Будући да је реч о ресурсу развијеном за енглески језик, његова примена није непосредна и не заснива се на пуком преводу, већ на адаптацији и темељној валидацији. То је нарочито важно у случајевима у којима не постоји потпуна лексичко-семантичка подударност између енглеског и српског језика. Због тога Принстонски ворднет у овом раду има двојаку функцију: он је истовремено полазни извор структурираног знања и ресурс чији се садржај критички преиспитује при његовом преношењу у српско језичко окружење.

4.2 Српски ворднет

Српски ворднет у овом раду представља централни лексичко-семантички ресурс који служи као основа за изградњу, валидацију и проширење скупа значења за српски језик. За разлику од Принстонског ворднета, који се користи као полазни и посреднички ресурс, Српски ворднет има средишњу улогу у експерименталном поступку, јер се резултати истраживања интегришу

управо у његов структурни оквир (Krstev et al. 2004; Obradović et al. 2004; Stanković, Mladenović, et al. 2018).

Српски ворднет је развијен као вишејезични и међујезични ресурс, заснован на структури Принстонског ворднета, при чему су синсети повезани са одговарајућим енглеским еквивалентима, а преко њега и са свим другим ворднетима који следе исту парадигму. Ова повезаност омогућава адаптацију значења, али показује и ограничења у њиховој покривености, глосама и семантичким релацијама. (Krstev et al. 2004). Управо ова ограничења представљају један од основних мотива за даље проширење ресурса у оквиру овог истраживања.

У практичном смислу, Српски ворднет се у овом раду користи и као улазни и као излазни ресурс. Као улаз, он обезбеђује постојећи инвентар значења, заједно са доступним глосама, литералима и релацијама, који се користе као основа за анализу и поређење. Као излаз, он представља циљну структуру у коју се интегришу новодобијена значења, проширене дефиниције и кориговани или новододати литерали. На тај начин, процес проширења непосредно се уграђује у постојећу лексичко-семантичку мрежу.

Посебан аспект употребе Српског ворднета у овом раду односи се на улогу глоса и семантичких релација у задацима који превазилазе чисто лексичку организацију, као што је анализа сентимента. Глосе, као експлицитне дефиниције значења, могу садржати вредносне и емоционалне компоненте релевантне за одређивање оријентације сентимента. Поред тога, поједине семантичке релације, нарочито антонимијске, могу бити значајне за моделовање супротности поларитета. Сличан принцип присутан је и у ресурсима као што је SentiWordNet, у којима се вредности сентимента додељују на нивоу синсета, узимајући у обзир структуру и односе унутар ворднета.

У том контексту, Српски ворднет у овом истраживању има оперативну улогу која превазилази функцију референтног ресурса. Његова двострука употреба, као извора података и као циљне структуре за проширење, омогућава да се резултати истраживања непосредно рефлектују на унапређење самог ресурса, чиме се доприноси његовој даљој примени у задацима обраде природног језика за српски језик (Stanković, Mladenović, et al. 2018).

4.3 ELEXIS-WSD и ELEXIS-WSD-sr

ELEXIS-WSD је вишејезички паралелни корпус намењен разрешењу вишезначности речи, развијен као ресурс у коме су реченице из више језика повезане са језички специфичним инвентарима значења. Прва верзија са десет европских језика (одељак 2.11) је базирана на енглеском скупу реченица из ресурса WikiMatrix, уз превођење, поравнавање и ручну проверу. Енглески део корпуса имао је организациону улогу: полазне енглеске реченице су филтриране према дужини и броју вишезначних речи, а затим су повезиване са одговарајућим преводима у другим језицима; у случајевима у којима одговарајући превод није био доступан, реченице су ручно превођене. Коначна поставка корпуса садржала је по 2.024 реченице за сваки језик (Martelli et al. 2021, 2023).

Сваки језик у оквиру ELEXIS-WSD има сопствени инвентар значења, прилагођен доступним лексичко-семантичким ресурсима датог језика. Инвентари садрже лему, универзалну врсту речи у складу са пројектом Universal dependencies (de Marneffe et al. 2021), идентификатор значења и дефиницију која се користи у поступку анотације. Њихови извори нису исти за све језике: за поједине језике засновани су на ворднетима тих језика, док су за друге језике изведени из националних дигиталних речника или других лексикографских извора. Оваква организација омогућава да се паралелни корпус користи у вишејезичким експериментима

разрешења вишезначности речи, али истовремено показује да је за сваки нови језик неопходно прилагодити инвентар значења расположивим локалним ресурсима (Martelli et al. 2021, 2023).

У наставку рада назив ELEXIS-WSD користи се за шири вишејезички ресурс, док се ELEXIS-WSD-sr односи на српски поткорпус и његову радну обраду у оквиру ове дисертације. Српски поткорпус ELEXIS-WSD-sr развија се у оквиру ширег проширења овог ресурса на друге језике у склопу COST акције CA21167 - Universality, diversity and idiosyncrasy in language technology (UniDive). У том контексту, српски, хрватски, румунски и други поткорпуси посматрају се као део даљег проширивања ELEXIS-WSD-а у оквиру активности усмерених на повећање језичке покривености, унапређење анотационих слојева и развој упоредивих смерница за вишејезичку семантичку анотацију (Čibej et al. 2025). За српски језик тај задатак има посебну методолошку тежину, јер не постоји јавно доступан, ручно означен WSD корпус који би могао непосредно да се употреби као полазни ресурс.

Полазиште за српски део рада чини Српски ворднет, који обезбеђује основну лексичко-семантичку структуру и везу са Принстонским ворднетом преко међујезичког индекса. Управо ова повезаност омогућава поређење и пренос значења између енглеског и српског ресурса. Међутим, постојећа покривеност Српског ворднета није довољна за све лексеме и значења која се јављају у корпусу, па је за потребе ове дисертације неопходно формирати радни инвентар значења. Тај инвентар се у првој фази ослања на енглески ELEXIS-WSD-en инвентар значења и његово поравнавање са Српским ворднетом, док се даље допуне изводе поступцима описаним у методолошком (поглавље 3) и резултатском делу рада (поглавља 5 и 6).

Српски поткорпус се у овој фази не посматра као завршен ELEXIS-WSD ресурс, већ као корпус који је прошао фазу почетне обраде и анотирања и који обезбеђује улаз за разрешење вишезначности речи. За реченице је доступно више слојева анотације, као што су леме, врсте речи, полилексемске јединице (изрази) и именовани ентитети. Ови слојеви су важни зато што поступак WSD-а у овој дисертацији полази од конкретне појаве лексеме у реченици, њене леме и врсте речи, а затим из радног инвентара преузима скуп кандидата за избор значења. Српски део ELEXIS-WSD-а зато има улогу улазног корпуса за експерименте разрешења вишезначности речи, док се сама семантичка анотација формира поступцима развијеним у овом истраживању. Квантитативни преглед анотираних значења, разврстаних према типу јединице и категоријама врсте речи (UPOS), дат је у Табели 4.1. Колона “Анотације” приказује укупан број анотираних једночланих или вишечланих јединица, колона “Облици” даје број различитих облика речи, док у колони “Леме” се приказује број различитих лема.

Табела 4.1 Статистика анотираних значења по врстама речи и типу јединице; корпус обухвата 2.024 реченице.

Тип јединице	UPOS	Анотације	Облици	Леме
једночлана	NOUN	6 513	3 776	2 233
	VERB	2 492	1 491	761
	ADJ	3 358	2 352	1 331
	ADV	988	302	254
једночлана	Укупно	13 351	7 921	4 579
вишечлана	NOUN	727	661	582
	VERB	428	333	236
	ADJ	5	5	5
	ADV	82	48	43
вишечлана	Укупно	1 242	1 047	866

Посебан део припреме српског поткорпуса односи се на вишечлане изразе и именоване ентитете. Претходни радови показују да су ови слојеви посебно важни за ELEXIS-WSD, јер се значење вишечланих израза често не може поуздано извести из значења њихових појединачних компоненти. Именички вишечлани изрази обрађивани су уз ослонац на електронске речнике (Krstev 2008), док су глаголски вишечлани изрази обрађивани комбинацијом електронских речника, алата Unitex (Paumier et al. 2009), правила и ресурса PARSEME. Именовани ентитети су означавани посебним моделима за српски језик, а затим повезивани са базом знања Википодаци (Vrandečić et al. 2023) и ручно проверавани и кориговани на платформи INCEpTION (Krstev, Stanković, and Marković 2024; Krstev, Stanković, Marković, et al. 2024).

У овој дисертацији ELEXIS-WSD-sr има две повезане улоге. Прво, он обезбеђује српске реченице са претходно припремљеним лингвистичким слојевима, што омогућава систематско издвајање циљних лексема за WSD. Друго, због паралелне структуре корпуса и постојања енглеског дела, ресурс омогућава циљану проверу употребе информација о значењу при превођењу са српског на енглески језик. У тој провери корпус се не користи за обучавање модела за превођење, већ као извор реченица у којима се може испитати да ли експлицитна информација о значењу помаже при избору енглеског еквивалента.

4.4 Речник савременог српског језика

Речник савременог српског језика (PCCJ) представља дигитални лексикографски ресурс у изради, речник који треба да опише речи које се често користе у савременом српском језику. Првенствено је намењен људима (различитих генерација и профила), али се предвиђа и његова примена у различитим рачунарским окружењима. У опису пројекта, PCCJ се представља као ресурс који треба да подржи претрагу и прегледање речничких података путем веб и мобилних апликација, али и програмски приступ преко API-ја. Пројекат развијају удружење „Окупљени око језика“ и Друштво за језичке ресурсе и технологије ЈеРТех, са циљем обраде приближно 50.000 фреквентних речи савременог српског језика. Лексички обухват заснива се на употреби у последњих педесет година и у различитим функционалним стиловима, што PCCJ разликује од ресурса који су превасходно настали дигитализацијом или преносом ранијих штампаних речника (Stanković et al. 2025).

Полазни списак одредница у PCCJ-у формиран је аутоматским издвајањем кандидата из великих корпуса савременог српског језика, пре свега *SrpKor2013* и *SrpKor2021* (Vitas et al. 2025), а затим се тај списак ручно проверава и допуњује лексикографским подацима из базе *LeXimirka* (Stanković, Krstev, et al. 2018). При избору и уређивању одредница узимају се у обзир фреквенција у корпусима, потврђеност у постојећим електронским и штампаним речницима, као и у Српском ворднету, али пресудна је процена и одлука лексикографа. На тај начин PCCJ комбинује корпусно засновану селекцију грађе са стручном лексикографском провером, што је важно због богатих варијантних облика, њиховог повезивања и потребе да се разликују савремена, стандардна, термилошка, разговорна и застарела употреба.

Технолошка инфраструктура PCCJ-а организована је као лексикографска база података у *PostgreSQL*-у, са моделом инспирисаним стандардом *DMLex* (Filip et al. 2025). Унутрашња структура речничког чланка обухвата одредницу, врсту речи, граматичке податке, значења, дефиниције, примере употребе, колокације, синтаксичке обрасце, синониме и упућивања на сродне јединице. За лексикографски рад развијена је прилагођена веб-апликација која повезује уређивање речничких чланака са корпусним претраживањем преко NoSke, локалне инстанце *NoSketchEngine*-а (Rychlý 2007; Kilgarrieff et al. 2014). Таква организација омогућава да се

лексикографска одлука не заснива само на изолованој дефиницији, већ и на корпусним примерима, колокацијама и подацима о фреквенцији употребе.

Посебан аспект РССЈ-а јесте употреба аутоматских поступака у различитим фазама израде речника. У раду (Stanković et al. 2025) су описани поступци за предлагање синонима помоћу векторских модела као што су *Word2Vec*, *FastText* и *Dict2Vec*, као и експерименти са великим језичким моделима у генерисању нацрта речничких дефиниција. У истом раду истиче се да ови поступци могу убрзати поједине фазе лексикографског рада, али да не могу заменити стручну редактуру. Аутоматски добијене дефиниције и примери нарочито су проблематични код вишезначних лексема, фразеолошких употреба, правописно осетљивих облика³⁰ и значења која зависе од ширег језичког и културног контекста. Због тога је РССЈ у овој дисертацији потребно посматрати као радни лексикографски ресурс у развоју, а не као коначан и потпуно валидиран речник.

За потребе ове дисертације РССЈ је коришћен као извор речничких значења и дефиниција у поступку лексичко-семантичког поравнања са Српским ворднетом. У том поступку речничко значење из РССЈ-а и синсет из Српског ворднета третирају се као јединице два различито организована лексичко-семантичка ресурса: РССЈ је организован око речничких одредница и њихових значења, док је Српски ворднет организован око појмова (концепата) представљених синсетима и семантичким релацијама. Издвојени узорак обухвата 535 речничких одредница и 1.829 речничких значења, која су у експерименту третирана као лексичка значења за поређење са синсетима Српског ворднета. Оваква поставка омогућава да се провери у којој мери се речничке дефиниције могу повезати са постојећом структуром организованом око синсета, као и где се јављају разлике у обухвату, степену општости и начину лексикографског описивања значења.

4.5 Senti-pol-sr

Лексикон *Senti-pol-sr* представља српски ресурс за анотацију сентимената, развијен интеграцијом три лексикона која су преведена са енглеских верзија — NRC, AFINN и Bing — уз накнадне додатне опсежне ручне корекције и допуне (Stanković et al. 2022). У поступку израде најпре су спојене одреднице из полазних лексикона, уклоњени дупликати и уједначени поларитети, а затим су елиминисане јединице које не припадају српском језику, на основу провере у српским морфолошким речницима. У случајевима у којима је унос био дат у флективном облику, лема је коригована, а новим лексичким јединицама додељене су и ознаке врсте речи. Завршну проверу ресурса ручном евалуацијом су урадила два анотатора (Stanković et al. 2022). За разлику од ворднета, у којима је организација заснована на синсетима, *Senti-pol-sr* је организован на нивоу лексичких јединица. Због тога овај лексикон у овом раду нема улогу централног извора за моделовање сентимента на нивоу значења, већ помоћног ресурса за избор и интерпретацију релевантних јединица. Његова конкретна улога у формирању узорка за анотацију описана је у методолошком делу рада, у одељку 3.2.

³⁰ Случајеви када правописна грешка може на први поглед да изгледа исправно али је у ствари са погрешним значењем. На пример, случајеви у којима је састављено или растављено писање повезано с разликом у значењу, као код „напамет“ и „на памет“

4.6 Коришћени велики језички модели

У овом одељку биће представљени велики језички модели и сродни модели векторске репрезентације који су коришћени или је њихово коришћење разматрано у експерименталном окружењу дисертације. Подела је изведена према архитектури и начину употребе модела, а не према појединачном експерименту: енкодерски модели користе се пре свега за добијање контекстуалних и реченичних репрезентација, декодерски модели за генерисање текста, док се затворени комерцијални генеративни и разговорни модели користе преко сервисног или програмског интерфејса. Називи модела и инвентара наведени су у изворном облику. Табела 4.2 приказује сажети преглед обухваћених модела.

У прву групу спадају **енкодерски модели**, односно модели који улазни текст претварају у контекстуалне или реченичне векторске репрезентације. *BERTić*³¹ је модел заснован на ELECTRA приступу и обучен на корпусу босанског, хрватског, црногорског и српског језика, што га чини релевантним за задатке обраде српског и блиских јужнословенских језика (Ljubešić and Lauc 2021). *Jerteh-355*³² је *RoBERTa-large* модел за српски језик, док је *TeslaXLM*³³ адаптација XLM-R-large архитектуре за српски и српскохрватски језички простор; оба модела припадају линији локално развијених или прилагођених енкодерских модела (Škorić and Petalinkar 2025; Škorić 2025). У овој групи су и *all-MiniLM-L6-v2*³⁴ и *intfloat/multilingual-e5-large*³⁵, који се користе као модели за реченичне репрезентације, семантичку сличност и претраживање по значењу (Reimers and Gurevych 2019; W. Wang et al. 2020; L. Wang et al. 2024).

Другу групу чине **декодерски, ауторегресивни модели** намењени генерисању текста. *GPT2-Orao*³⁶ је локални генеративни модел за српски језик, заснован на GPT-2-large архитектури и обучен на приближно четири милијарде токена, који подржава ћирилицу и латиницу (Škorić 2025). *Mistral 7B Instruct*³⁷ и *Mistral-Small-3.2-24B-Instruct-2506*³⁸ представљају отворене моделе дообучене инструкцијама из породице Mistral, погодне за контролисано генерисање, рад са промптовима и поређење отворених генеративних модела различите величине (Jiang et al. 2023; Mistral AI 2025). *Meta Llama 4 Scout*³⁹ је модел дообучен инструкцијама са отвореним тежинама, заснован на архитектури мешавине експерата, са јавном картицом модела која наводи 17 милијарди активних и 109 милијарди укупних параметара (Meta 2025).

Трећу групу чине **затворени комерцијални генеративни и разговорни модели**, пре свега верзије *ChatGPT/GPT* и *Gemini 2.5 Flash* и *Gemini 2.5 Pro*. За ове моделе нису јавно доступни сви технички детаљи, па се у дисертацији број параметара наводи само ако је потврђен у званичној документацији. Иако поједине верзије ових система подржавају и мултимодални рад, у овој дисертацији оне се посматрају превасходно као текстуални генеративни модели. Зато се описују преко породице модела, доступности, контекстног прозора када је он документован и улоге у експериментима, као што су разрешење вишезначности речи, превођење, генерисање дефиниција, предлагање примера употребе и редакција излаза (OpenAI 2023; Comanici et al. 2025; Gemini Team et al. 2025). Оваква организација омогућава да се модели пореде по архитектури и доступности, док се њихова конкретна експериментална улога

³¹ <https://huggingface.co/classla/bcml-bertic>

³² <https://huggingface.co/jerteh/Jerteh-355>

³³ <https://huggingface.co/te-sla/TeslaXLM>

³⁴ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

³⁵ <https://huggingface.co/intfloat/multilingual-e5-large>

³⁶ <https://huggingface.co/jerteh/gpt2-orao>

³⁷ <https://ollama.com/library/mistral:7b-instruct>

³⁸ <https://ollama.com/library/mistral-small3.2:24b>

³⁹ <https://ollama.com/library/llama4:scout>

издваја као посебна димензија. Због тога се *TeslaXLM*, на пример, не своди на један задатак, већ се описује као XLM-R модел користан за више типова векторизације и поређења текста.

У циљаној провери употребе слоја за разрешење вишезначности речи у машинском преводу коришћени су постојећи **системи за превод** са српског на енглески језик, без додатног обучавања на материјалу ове дисертације. Ови системи нису посматрани као предмет опште евалуације квалитета машинског преводу, већ као различита преводилачка окружења у којима се проверава да ли кратка информација о значењу може да утиче на избор енглеског преводног еквивалента.

Основна комерцијална поставка, у резултатима означена као *Google Basic*, односи се на *Google Cloud Translation - Basic API (v2)*⁴⁰, односно на *REST translate* метод сервиса *Cloud Translation*. Ова поставка подразумевамо користи *Google Neural Machine Translation (NMT)*, без прилагођеног преводилачког модела. Ознака *Google Basic* у овом раду зато не означава посебан модел развијен у оквиру дисертације, већ основну API поставку сервиса *Google Cloud Translation* која је коришћена као практично доступна полазна тачка за проверу преводу.

Поред ове поставке, проверени су и *OPUS-MT/Marian*⁴¹ и *M2M-100*⁴². *OPUS-MT* је пројекат отворених ресурса и претходно обучених модела за машинско преводу (Tiedemann and Thottingal 2020), док је *Marian* оквир за неуронско машинско преводу, написан у C++ и намењен ефикасном обучавању модела за преводу, укључујући и закључивање (енгл. reasoning) (Junczys-Dowmunt et al. 2018). У овој дисертацији назив *OPUS-MT/Marian* означава употребу отвореног преводилачког модела из *OPUS-MT* окружења преко *Marian* инфраструктуре.

M2M-100 је вишејезички модел за машинско преводу заснован на поставци преводу између више изворних и више циљних језика (енгл. *many-to-many*), односно на директном преводу између великог броја језичких парова у оквиру једног модела (Fan et al. 2021). Његова улога у овој провери била је да се испита да ли вишејезички преводилачки модел другачије реагује на уметнуте информације о значењу него основна комерцијална API поставка и *OPUS-MT/Marian* окружење.

Табела 4.2 Преглед коришћених великих језичких и репрезентационих модела

Модел	Тип/ архитектура	Породица/ос нова	Парамет ри	Језици	Доступност
BERTiс	енкодерски	ELECTRA; classla/bcms- bertic	110M	BCMS	отворене тежине ⁴³
Jerteh-355	енкодерски	RoBERTa- large; jerteh/Jerteh- 355	355M	српски; ћир. и лат.	отворене тежине
TeslaXLM	енкодерски	XLM-R-large; te- sla/TeslaXLM	561M	српски/српскохр в.; ћир. и лат.	отворене тежине

⁴⁰ <https://docs.cloud.google.com/translate/docs/reference/rest/v2/translate>

⁴¹ <https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-sh-en>

⁴² https://huggingface.co/facebook/m2m100_418M

⁴³ Модел са отвореним тежинама (open-weight model) јесте модел чији су параметри (тежине) јавно доступни за преузимање и локално коришћење, у складу са условима одговарајуће лиценце.

all-MiniLM-L6-v2	реченични енкодер	MiniLM-L6; sentence-transformers	22,7М	превасходно енглески	отворене тежине
multilingual-e5-large	вишејезички реч. енкодер	XLNet-R-large; E5	око 0,6В	вишејезички; 94/100 језика	отворене тежине
GPT2-Orao	декодерски	GPT-2-large; jerteh/gpt2-oraо	800М	српски; ћир. и лат.	отворене тежине
Mistral 7B Instruct	инструктивн и декодер	Mistral 7B	7В	општи вишејезички	отворене тежине
Mistral Small 3.2 24В	инструктивн и декодер	Mistral Small 3.2	24В	вишејезички; 24 језика	отворене тежине
Meta Llama 4 Scout	МоЕ декодер	Llama 4 Scout 17Вх16Е	17В активних ; 109В укупно	вишејезички; 12 језика	отворене тежине
ChatGPT/GPT	затворени ген./разговор ни	GPT породица	није јавно наведено	вишејезички	сервис/API
Gemini 2.5	затворени ген./разговор ни	Gemini 2.5	није јавно наведено	вишејезички	сервис/API
Cloud Translation Basic (v2)	Cloud MT сервис	Google NMT; Cloud Translation	није јавно наведено	100+ језика	Cloud API
OPUS-MT/Marian	отворени NMT	OPUS-MT; Marian	око 0,2В	sr-en	отворено; локално
M2M-100	вишејезички MT	Transformer; many-to-many	418М	100 језика	отворене тежине

4.7 Скуп новинских реченица за евалуацију сентимента

За екстерну евалуацију лексикона сентимента припремљен је скуп реченица из новинских текстова на српском језику, са ручно додељеним ознакама поларитета. Скуп је намењен провери употребљивости вредности сентимента у аутоматском одређивању поларитета реченице. Јединица ручне анотације јесте реченица као целина, а њена ознака служи као референтна вредност са којом се пореди резултат примене лексикона. Тиме се евалуација на нивоу синсетова допуњује провером ресурса у обради текста.

Полазни узорак обухвата 3.000 реченица из 3.000 различитих чланака. Реченице су насумично издвојене уз уједначавање узорка према дужини и ограничење на једну реченицу по чланку. Издвојено је по 1.000 кратких, средњих и дугих реченица, са 8–14, 15–24 и 25–45 речи. Ове категорије односе се на дужину реченица, док је расподела према поларитету утврђена накнадним ручним означавањем.

Узорак је распоређен међу 13 анотатора, са по 230 или 231 реченицом, тако да је свака реченица додељена једном анотатору. За означавање су коришћене табеле са писаним упутством, текстом реченице, пољем за ознаку сентимента и пољем за необавезну напомену. Анотатори су бирали једну од три вредности: –1 за негативан, 0 за неутралан и 1 за позитиван сентимент.

Напомене су биле предвиђене за случајеве ироније, нејасног контекста, више могућих тумачења или проблема у тексту.

Према упутству, негативна ознака додељивала се реченицама које изражавају негативан став, критику, проблем, штету, опасност или лош исход, а позитивна реченицама које изражавају позитиван став, похвалу, успех, корист, побољшање или добар исход. Неутрална ознака била је предвиђена за информативне и описне реченице, као и за случајеве у којима сентимент није јасан без ширег контекста. Зато неутрална класа у овом скупу обухвата и реченице код којих је процена поларитета ограничена доступним контекстом, што треба узети у обзир при тумачењу резултата.

Прикупљене табеле 11 анотатора садрже 2.539 реченица, од којих три немају додељену ознаку. После њиховог изостављања, коначни евалуациони подскуп обухвата 2.536 реченица. Преостала 461 реченица из почетне расподеле није обухваћена прикупљеним табелама. Расподела коначног подскупа према поларитету приказана је у следећој табели.

Расположиви подаци садрже по једну ручну ознаку за сваку укључену реченицу, па на основу њих није могуће проценити међуанотаторску сагласност. Ручне ознаке стога представљају референтне процене појединачних анотатора, чија међусобна сагласност на истим примерима није утврђена. Додатно ограничење проистиче из избора новинске грађе: добијени резултати описују успешност на овом скупу, а њихова применљивост на друге врсте текста захтева засебну проверу.

5 Резултати

У овом поглављу резултати истраживања ће бити приказани кроз четири међусобно повезане целине: развој инфраструктуре за рад са Српским ворднетом, изградња и евалуација лексикона сентимента, изградња, проширење и евалуација инвентара значења за разрешење вишезначности речи, и провера примене тако обogaћених ресурса у одабраним практичним низводним задацима (енгл. *downstream tasks*).

5.1 Омотач за ворднет

За потребе овог истраживања било је неопходно да се развије интерфејс који машинама омогућава систематску обраду синсета. Циљ није био само да се поново реализује постојећи функционални омотач већ и да се омогуће програмско формирање и проширивање скупова поларитета, припрема скупова података за обучавање и примена модела машинског учења. Интерфејс је, стога, конципиран као експериментална инфраструктура за изградњу и евалуацију лексикона сентимента.

Полазну основу за изградњу овог интерфејса представља XML репрезентација целокупне верзије Српског ворднета или неког његовог подскупа. Развијен је програм који омогућава учитавање целокупне структуре у радну меморију, што је изводљиво имајући у виду релативно ограничен обим ресурса. Оваква организација омогућава ефикасно претраживање, навигацију кроз семантичке релације и модификацију синсета.

Посебна пажња посвећена је моделовању синсета и лема на начин који је компатибилан са постојећим NLTK структурама, уз уважавање специфичности *SrpWN*-а. За разлику од *PWN*-а, где је идентификација синсета заснована на имену леме и редном броју значења (нпр. *salt.n.01* је идентификатор синсета **ENG30-05717342-n** који садржи литерал *salt* у првом значењу), у *SrpWN*-у синсет је означен алфанумеричком идентификационом шифром, односно интерлингвалним индексом *ILI* (нпр. *ENG30-14672023-n*). Поред тога, синсети у коришћеној верзији ресурса садрже и експлицитно придружене вредности сентимента. Интерфејс је прилагођен овим структурним разликама, уз задржавање функционалне компатибилности са NLTK приступом.

Пакет је организован око два повезана нивоа. Први ниво чини инфраструктурни модул за читање и представљање Српског ворднета. У њему су кључне класе *SrbWordNetReader*, *SrbSynset* и *SrbLemma*. Класа *SrbWordNetReader* проширује начин рада NLTK читача ворднета и обезбеђује методе за проналажење синсета по идентификатору (*synsetfromID*), по леми и врсти речи (*synsets*), по комбинацији врсте речи и офсета⁴⁴ (*synsetfromposandoffset*), као и за учитавање помоћног морфолошког лексикона (*loadlexicon*) и једноставну лематизацију облика (*morph*). Класа *SrbSynset* представља појединачни синсет и обезбеђује приступ идентификатору, дефиницији, примерима употребе, лемама, врсти речи, релацијама антонимије, деривационо повезаним облицима и осталим релацијама које су дефинисане у ворднету. Класа *SrbLemma* представља појединачни литерал унутар синсета, заједно са ознаком значења и напоменама које су садржане у XML структури.

⁴⁴ Офсет у идентификатору синсета означава позицију његовог записа, изражену бројем бајтова од почетка одговарајуће датотеке Princeton WordNet-а. Заједно са ознаком врсте речи, офсет једнозначно одређује синсет у оквиру одређене верзије овог ресурса. Овај начин идентификације примењен је и у представљању Princeton WordNet-а у библиотеци NLTK.

Други ниво чини експериментални слој за рад са информацијом о сентименту. Његово језгро је класа *PolaritySets*, која представља програмску апстракцију позитивног, негативног и објективног скупа синсета. Ова класа омогућава ручно додавање синсета или њихових идентификатора у одговарајуће скупове (*addPOS*, *addNEG*, *addPOSID*, *addNEGID*), учитавање полазних ознака из ресурса *Micro-WNOp* и *SentiWordNet* (*addWNopAll*, *addWSWNALL*), као и итеративно проширивање скупова преко семантичких релација (*next_iteration*). Функција *getXY* на основу података о синсетима и њиховој припадности наведеним скуповима припрема улазне податке и одговарајуће ознаке за обучавање класификатора. Помоћна класа *SrbSynset2GlossTransformer* из одабраних синсета издваја њихове глосе, које се затим помоћу алата библиотеке *scikit-learn* могу претворити у нумерички облик погодан за обучавање класификатора.

Имплементација омогућава:

- програмску експанзију скупова поларитета сентимента на основу семантичких релација,
- формирање табеларних репрезентација синсета и њихових својстава,
- аутоматско генерисање скупова података за обучавање, погодних за примену метода машинског учења,
- интеграцију са класичним и моделима за класификацију заснованим на неуронским мрежама.

Посебно је омогућено издвајање парова улаз–ознака (X, y), где се синсет репрезентује преко своје глосе или других лексичко-семантичких својстава, док се ознаке формирају на основу скупова поларитета сентимента. На тај начин интерфејс представља инфраструктурни слој који повезује симболичку семантичку мрежу са статистичким и моделима заснованим на неуронским мрежама.

Архитектура интерфејса раздваја инфраструктурни слој (учитавање и структурирање XML података) од експерименталног слоја (експанзија вредности поларитета, векторизација, класификација и евалуација). Оваква организација омогућава репродуктивност експеримената, контролу параметара експанзије и систематско испитивање различитих варијанти методологије. Развијени пакет је јавно доступан⁴⁵, чиме се омогућава његова употреба у другим истраживањима и интеграција *SrpWN*-а у постојеће системе за обраду природног језика који су имплементирани у језику Python.

5.2 Евалуација модела обучених за процену сентимента

Евалуација модела спроведена је применом стандардних мера учинка класификације: прецизности, одзива и F1 мере. Све метрике израчунате су као макро-просеци, како би се обезбедило једнако третирање обе класе: позитивне и негативне (P_k и N_k), независно од њихове заступљености у скупу података. Овај приступ је посебно релевантан у условима

⁴⁵ Python омотач за Српски ворднет доступан је у GitHub репозиторијуму:
<https://github.com/sasa5linkar/Srpski-Wordnet-Python-wrapper>

Пакет је доступан и преко PyPI:
<https://pypi.org/project/srpskiwordnet/>

Инсталација:
`pip install srpskiwordnet`

потенцијалне небалансираности класа, јер спречава доминацију већинске класе у укупној оцени учинка.

Модели су тестирани на четири скупа података (T0, T2, T4 и T6)⁴⁶, са циљем процене стабилности и робусности у различитим конфигурацијама података. Табела 5.1 приказује детаљне резултате, док су кључни налази сумирани у наставку.

Табела 5.1 Метрике класификације добијене на евалуационом скупу за моделе AdaBoost, NB, SVM, RNN, Transformer, BERTiћ, GPT2-Orao и Jerteh-355

Model		T0			T2			T4			T6		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
AdaBoost	Pk	0,67	0,59	0,62	0,65	0,62	0,63	0,61	0,57	0,59	0,60	0,58	0,59
	Nk	0,80	0,67	0,71	0,69	0,63	0,65	0,68	0,59	0,62	0,70	0,63	0,66
NB	Pk	0,53	0,67	0,54	0,55	0,68	0,57	0,53	0,61	0,55	0,53	0,61	0,55
	Nk	0,55	0,71	0,58	0,55	0,65	0,57	0,55	0,65	0,57	0,56	0,67	0,58
SVM	Pk	0,53	0,81	0,54	0,53	0,84	0,51	0,55	0,78	0,57	0,53	0,85	0,48
	Nk	0,51	0,75	0,39	0,51	0,66	0,28	0,52	0,74	0,37	0,52	0,73	0,36
RNN	Pk	0,50	0,50	0,50	0,79	0,55	0,59	0,89	0,53	0,55	0,49	0,50	0,50
	Nk	0,87	0,56	0,59	0,89	0,53	0,55	0,67	0,54	0,56	0,99	0,51	0,52
Transformer	Pk	0,67	0,59	0,68	0,75	0,71	0,73	0,57	0,66	0,59	0,67	0,67	0,67
	Nk	0,75	0,71	0,73	0,69	0,67	0,68	0,67	0,68	0,67	0,62	0,70	0,65
BERTiћ	Pk	0,79	0,89	0,83	0,87	0,68	0,74	0,85	0,69	0,75	0,89	0,66	0,73
	Nk	0,89	0,84	0,86	0,82	0,88	0,85	0,91	0,87	0,89	0,84	0,86	0,85
GPT2-Orao	Pk	0,78	0,77	0,77	0,66	0,61	0,63	0,80	0,55	0,59	0,61	0,62	0,62
	Nk	0,76	0,72	0,77	0,75	0,61	0,65	0,67	0,73	0,70	0,77	0,65	0,69
Jerteh-355	Pk	0,87	0,84	0,85	0,99	0,59	0,65	0,92	0,72	0,79	0,81	0,72	0,76
	Nk	0,87	0,68	0,74	0,82	0,79	0,80	0,87	0,66	0,72	0,93	0,65	0,72

Напомена: P означава прецизност, R одзив, а F1 хармонијску средину прецизности и одзива. Термин „Transformer“ у табели означава компактну трансформерску архитектуру обучену од почетка на посматраном скупу података, док су модели BERTiћ, GPT2-Orao и Jerteh-355 претходно тренирани велики језички модели који су у оквиру овог рада накнадно фино подешени за задатак класификације сентимента.

Резултати показују да AdaBoost испољава стабилан учинак кроз све конфигурације, уз благо опадање вредности метрика. У поређењу са њим, модел заснован на наивном Бајесовом класификатору (NB) демонстрира конзистентан, али умерен учинак, са мањим осцилацијама између различитих скупова података. С друге стране, модел заснован на алгоритму потпорних вектора (SVM), са Гаусовим језгром (RBF), систематски постиже слабије резултате. Иако остварује релативно висок одзив, његова прецизност је знатно нижа што указује на изражену склоност ка лажно позитивним предвиђањима. Упркос релативно високом одзиву, ниска прецизност доводи до ниске вредности F1 мере.

Поређење рекурентне неуронске мреже (RNN) и „лаког“ трансформерског модела обученог од почетка показује да RNN у појединим конфигурацијама постиже већу прецизност, док трансформер надмашује RNN према обухвату и F1 мери. Будући да је задатак сведен на два бинарна класификатора, позитивни и негативни примери не означавају супротстављене сентименте. Један класификатор утврђује присуство или одсуство позитивног сентимента, а други присуство или одсуство негативног сентимента. Резултати стога указују на бољу

⁴⁶ Ознаке T0, T2, T4, T6 односе се на итеративно формиране скупове података за обучавање, при чему индекс означава редни број итерације. Скуп T0 представља почетни, ручно припремљен и означен скуп, док су скупови T2, T4, T6 формирану у наредним итерацијама поступка описаног у одељку 3.1.

способност трансформерске архитектуре да у оба случаја успостави равнотежу између прецизности препознавања сентимента и обухвата примера у којима је он присутан. У групи претходно тренираних модела који су додатно фино подешавани за конкретни задатак (BERTić, GPT2-Orao и Jerteh-355), уочавају се значајне разлике у стабилности и укупном учинку. BERTić показује конзистентно високе и уравнотежене вредности прецизности, обухвата и F1 мере у свим конфигурацијама, што указује на стабилан и поуздан учинак. Jerteh-355 такође демонстрира снажне и релативно стабилне резултате, посебно у погледу прецизности, и у већини случајева постиже вредности упоредиве са BERTić моделом.

Насупрот томе, GPT2-Orao испољава израженију варијабилност, са значајним осцилацијама у вредностима прецизности и F1 мере, што указује на мању стабилност у односу на друга два ВЈМ модела. Да би се формално проценила статистичка значајност разлика између модела, примењен је Фридманов тест над F1 вредностима кроз све итерације и скупове ознака, уз накнадну Nemenyi post-hoc анализу. Фридманов тест дао је вредност статистике $\chi^2 = 50,95$ ($p < 0,00000001$), што указује на статистички значајне разлике у учинку модела.

Највише просечне рангове остварили су:

- BERTić (1,38),
- Jerteh-355 (1,88),

при чему се BERTić издваја као најуспешнији модел. GPT2-Orao (3,50) и трансформер (3,63) заузимају средишњу позицију, без значајне међусобне разлике. AdaBoost (4,63) показује умерен учинак, док NB (6,50), RNN (6,94) и SVM (7,56) заузимају ниже рангове.

5.2.1 Евалуација добијених лексикона сентимента

Евалуација генерисаних лексикона у идеалним условима подразумевала би њихову примену на корпус са ручно анотираним сентиментом, уз употребу алата способног да повезује лексеме у тексту са одговарајућим синсетима. Алтернативно, могућа је процена на мањем скупу синсета са ручно додељеним вредностима поларитета. Међутим, у тренутку спровођења истраживања такви ресурси за српски језик нису били доступни, те је усвојен индиректни приступ евалуацији.

У оквиру тог приступа примењена су два комплементарна поступка. Први поступак заснива се на ручној евалуацији синсета којима су генерисани лексикони доделили највише вредности позитивног или негативног поларитета. Тај поступак омогућава проверу оних случајева у којима лексикони испољавају највећу сигурност и непосредно показује да ли високе вредности поларитета одговарају процени коју су приписали људи. Други поступак заснива се на поређењу лексикона са посебно формираним синтетичким скупом за евалуацију, добијеним уз подршку генеративног модела и накнадно провереним на ограниченом узорку. На тај начин се резултати не процењују само кроз екстремне вредности поларитета, већ и у односу на независно формирану процену оријентације сентимента синсета.

5.2.2 Ручна евалуација добијених лексикона сентимента

Лексикони су анализирани на основу синсета са највишим вредностима поларитета, као и на основу дистрибуције синсета по категоријама поларитета. Ручна евалуација спроведена је над 279 синсета од стране три независна оцењивача, а коначна одлука донета је већинским гласањем (Petalinkar et al. 2025). Сагласност између анотатора процењена је применом

Флајсове к мере, при чему је добијена вредност 0,8025, што указује на значајан степен међусобне сагласности.

За ручну евалуацију издвојени су синсети са вредношћу позитивног или негативног поларитета већом од 0,9 у најмање једном лексикону, означени као „снажно позитивни“, односно „снажно негативни“. Аотатори су оцењивали да ли је предложена класификација исправна или погрешна. На основу описаног већинског правила потврђена је позитивна класификација 71 синсета и негативна класификација 168 синсета.

Резултати ручне евалуације конструисаних лексикона приказани су у Табели 5.2. У табели је наведен укупан број синсета које је сваки лексикон идентификовао као снажно позитивне и снажно негативне, као и број исправно и неисправно идентификованих синсета у односу на ручну процену аотатора. Колона „Тачно“ означава синсете чија је класификација поларитета потврђена већинским ставом аотатора, док „Нетачно“ обухвата синсете који нису били правилно идентификовани у односу на ручну процену.

Табела 5.2 Резултати евалуације конструисаних лексикона сентимента и њихово поређење са иницијалним лексиконом S0

Лексикон	Позитивни			Негативни		
	Снажно	Тачно	Нетачно	Снажно	Тачно	Нетачно
S0	6	4	67	1	1	105
S1	1	1	70	26	18	88
S2	23	21	50	65	57	49
S4	55	48	23	106	98	8
S6	25	20	51	96	87	19

Лексикон S1 доноси побољшање у идентификацији негативних синсета (26 снажно негативних, од којих је 18 исправно класификовано), али и даље показује врло ограничену покривеност позитивне класе. Значајнији напредак уочава се код лексикона S2, који идентификује 23 снажно позитивна синсета (21 исправно) и 65 снажно негативних синсета (57 исправно), уз приметно смањење броја пропуштених случајева у односу на S0 и S1.

Најбоље резултате у оквиру позитивне класе постиже лексикон S4, са 55 снажно позитивних синсета, од којих је 48 потврђено као исправно класификовано, уз значајно смањен број пропуштених случајева (23). Истовремено, S4 бележи 106 снажно негативних синсета, од којих је 98 потврђено, што представља највиши степен покривености негативне класе међу свим анализираним верзијама.

Лексикон S6 такође показује висок степен учинка у негативној класи (96 снажно негативних, од којих је 87 исправно класификовано), али уз нешто мању покривеност позитивне класе у поређењу са S4. Лексикони S3, S5 и S7 нису укључени у табеларни приказ јер ниједан синсет у овим верзијама није достигао праг изразито високе поларитетске вредности ($> 0,9$). Сходно томе, анализа снажно позитивних и снажно негативних синсета за ове лексиконе није била применљива.

5.2.3 Формирање и провера скупа за евалуацију сентимента синсета

У поступку аотације модел је успешно обрадио свих 75 синсета из уравнотеженог узорка, формираног у складу са методологијом описаном у одељку 3.2. Ниједан пример није одбачен због невалидног излаза, а ни у једном случају није било потребно нормализовати збир позитивне и негативне компоненте. Сви синсети добили су нулту вредност позитивне или негативне компоненте, што показује да модел није додељивао поларитет тамо где није препознао јасну позитивну или негативну оријентацију. Расподела добијених вредности

приказана је у Табели 5.3. Табела раздваја позитивну и негативну компоненту, јер су оне у поступку аотације процењиване независно.

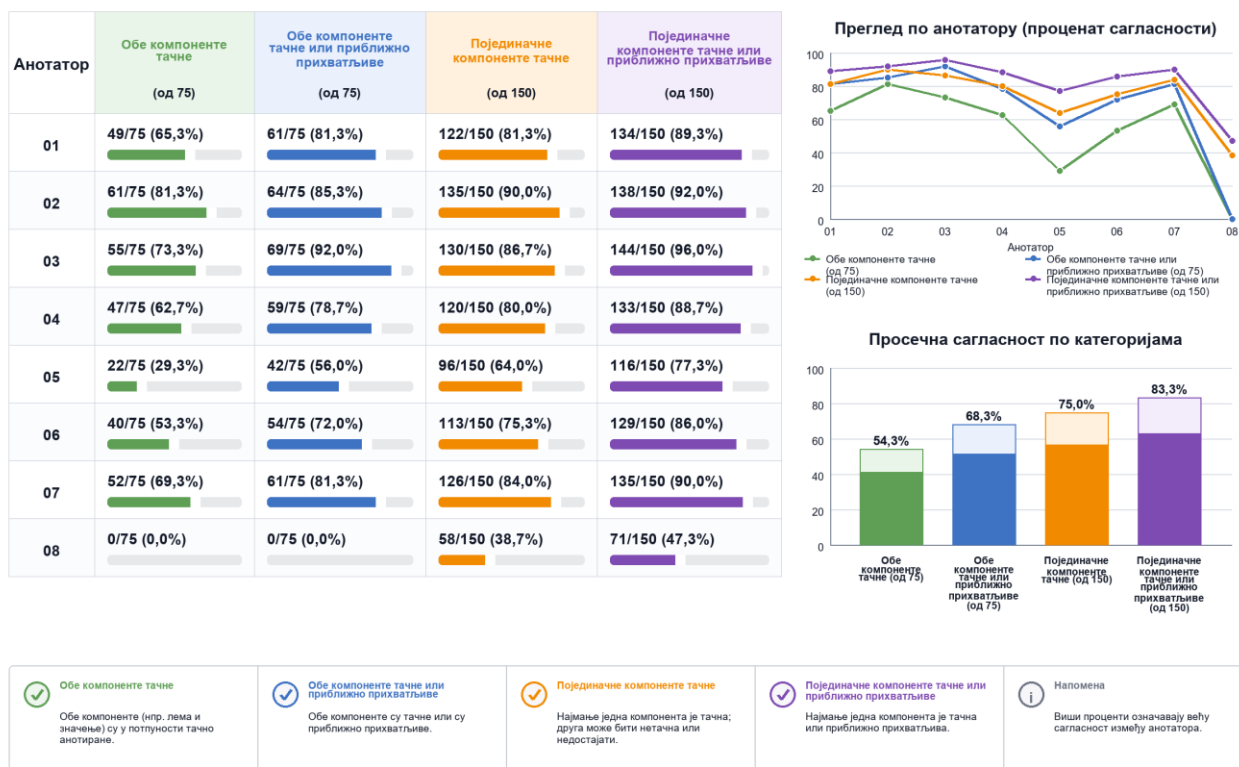
Табела 5.3 Расподела вредности позитивне и негативне компоненте у уравнотеженом узорку синсета

Вредност	Број синсета (POS)	Број синсета (NEG)
0	54	60
0,25	3	3
0,5	15	3
0,75	2	4
1	1	5
Укупно	75	75

У овом узорку нису забележени случајеви у којима су обе компоненте истовремено веће од нуле. То не представља ограничење самог поступка, јер је моделу експлицитно допуштено да додели и позитивну и негативну вредност истом синсету, у складу са структуром SentiWordNet-а. Међутим, у обрађеном узорку модел је све синсете расподелио тако да је најмање једна од две компоненте остала једнака нули. Овај налаз је користан за тумачење резултата, јер показује да раздвајање позитивног и негативног промпта није довело до механичког додељивања обе врсте поларитета истим примерима.

Након аутоматске доделе категоричких вредности спроведена је ручна провера добијеног синтетичког скупа за евалуацију. Провера је до сада обухватила осам независних аотатора, студената који су претходно добили упутство за поступак провере свих 75 синсета у уравнотеженом узорку. Аотатори нису самостално додељивали вредности сентимената синсетима, већ су проверавали вредности и образложења која је предложио ВЈМ. За сваку позитивну и негативну компоненту процењивали су да ли је вредност коју даје модел тачна, приближно прихватљива или нетачна; у случајевима неслагања могли су да предложе исправљену вредност. У једном од додатих образаца поједина поља за статус оцено остала су непопуњена јер један од аотатора није довршио аотацију. Због тога се резултати засновани на статусним ознакама за тај образац сматрају делимичним, док су у поређењу задржане коначне вредности, утврђене на основу унетих исправки и непромењених вредности које је доделио модел. Резултати по појединачним аотаторима приказани су на Слици 5.1.

Слика 5.1 показује да међу аотаторима постоје разлике у строгости, при чему је пети аотатор најстрожи међу обрасцима са доследно попуњеним статусима. Ред за осмог аотатора треба читати издвојено: нулте вредности у прве две колоне пре свега одражавају велики број непопуњених статусних поља, а не 75 експлицитно негативних оцена. У анализу су укључене све благовремено достављене аотације, будући да њихов број није био велики. Због могућих разлика у појединачним проценама, за даље тумачење резултата погодније је ослонити се на већинску одлуку и мере међуанотаторске сагласности него на оцену само једног аотатора.



Слика 5.1 Резултати ручне провере синтетичког скупа за евалуацију по анотатору

Резултати провере указују на то да је синтетички скуп у највећем броју случајева у складу са проценом људи. Посматрано по појединачним компонентама, већинска оцена осам анотатора је потпуно тачна у 123 од 150 случајева (82,0%). Ако се оцене „тачно“ и „близу“ посматрају као прихватљиви исходи, већински прихватљивих је 137 од 150 компоненти (91,3%). На нивоу целог синсета, обе компоненте су већински оцењене као потпуно тачне у 49 од 75 случајева, док су обе компоненте већински прихватљиве у 62 од 75 случајева. Детаљнији приказ сагласности дат је у Табели 5.4.

Табела 5.4 Сагласност анотатора у ручној провери синтетичког скупа за евалуацију

Мера сагласности	Обухват	Вредност
Потпуна сагласност о оцени вредности („тачно“, „близу“, „нетачно“)	све компоненте	24/150 (16,0%)
Постоји већинска одлука о оцени вредности	све компоненте	132/150 (88,0%)
Флајсова к мера за оцену вредности	све компоненте	0,1693
Потпуна сагласност о прихватљивости	све компоненте	47/150 (31,3%)
Постоји већинска одлука о прихватљивости	све компоненте	147/150 (98,0%)
Флајсова к мера за прихватљивост	све компоненте	0,1634
Потпуна сагласност о коначној позитивној вредности	позитивна компонента	23/75 (30,7%)
Све вредности позитивне компоненте у оквиру једног корака	позитивна компонента	38/75 (50,7%)

Просечна квадратно пондерисана к мера за парове анотатора	позитивна компонента	0,6885
Потпуна сагласност о коначној негативној вредности	негативна компонента	41/75 (54,7%)
Све вредности негативне компоненте у оквиру једног корака	негативна компонента	58/75 (77,3%)
Просечна квадратно пондерисана к мера за парове анотатора	негативна компонента	0,7757

Сагласност између анотатора показује да се овај скуп може користити као контролни скуп за евалуацију, али не и као некритички прихваћен аутоматски златни стандард. После укључивања седмог и осмог анотатора (који су анотирали касније) потпуна сагласност о статусу оцене износи 16,0% појединачних компоненти, док већинска одлука постоји у 88,0% случајева. Када се разликује само да ли је вредност прихватљива или неприхватљива, потпуна сагласност износи 31,3%, а већинска одлука постоји у 98,0% компоненти. Ниже вредности Флајсове к мере за статус оцене и за прихватљивост треба тумачити у светлу неуједначене расподеле класа, разлика у строгости анотатора и непотпуно попуњених статусних поља у једном обрасцу. За коначне степеноване вредности сентимента, просечна квадратно пондерисана к мера израчуната за парове анотатора износи 0,6885 за позитивну и 0,7757 за негативну компоненту, што показује да је сагласност стабилнија на негативној компоненти, док је позитивна компонента осетљивија на разлике у процени.

У односу на полазно пресликавање из SentiWordNet-a, сви синсети у овом узорку изабрани су управо зато што су имали неутралну вредност (0,0) у постојећем пресликавању, док је лексички ресурс Senti-pol-sr указивао на могућу вредност поларитета коју треба обележити, и то бар за један литерал. Ручна провера са осам анотатора показује да постоји висок степен прихватљивости вредности добијених моделом, али и да строга већинска коначна вредност није увек доступна, што је очекивано када се оцењује степен поларитета на дискретној скали. Због тога је за овај рад важније питање да ли провера коју су спровели људи налаже замену вредности добијених моделом. Тај однос је приказан у Табели 5.5.

Табела 5.5 Однос већинских коначних вредности анотатора и вредности добијених моделом

Обухват	Већинска коначна вредност постоји	Иста као вредност модела	Различита од вредности модела	Без строге већинске вредности
Позитивна компонента	66/75	64/66	2/66	9/75
Негативна компонента	63/75	61/63	2/63	12/75
Обе компоненте синсета	55/75	51/55	4/55	20/75

Табела 5.5 показује да ручна провера не даје разлог за систематску измену вредности које је модел доделио у синтетичком евалуационом скупу. Код позитивне компоненте већинска

коначна вредност постоји у 66 случајева, а у 64 случаја поклапа се са вредношћу коју је дао модел. Код негативне компоненте већинска коначна вредност постоји у 63 случаја, од чега се у 61 случају поклапа са вредношћу добијеном од модела. На нивоу синсета, обе већинске коначне вредности постоје у 55 случајева, а у 51 случају обе су исте као вредности које је дао модел. Поред тога, потпуна сагласност свих осам анотатора о истој корекцији у односу на вредност коју је дао не јавља се ни у једном синсету. Зато је методолошки оправдано да се у поређењу лексикона задрже изворне вредности које је дао модел синтетичког скупа за евалуацију, а да се резултати ручне провере наведу као потврда његове употребљивости. Овај налаз уједно потврђује да лексички сигнал поларитета није довољан за аутоматско додељивање сентимента синсету, али је користан као филтер за издвајање случајева у којима вреди проверити неутрално пресликавање из SWN.

У садашњем стању, подаци показују да је поступак аутоматског формирања синтетичког скупа за евалуацију изводљив и да добијени скуп садржи довољно велики број вредности прихватљивих људима да може да послужи за контролисану проверу лексикона сентимената. Зато се у наставку дисертације може говорити о синтетичком евалуационом скупу који је накнадно ручно проверен, али не и о потпуно независном златном стандарду. Додавање нових анотатора појачава опрез у тумачењу појединачних спорних синсета, али не мења основни закључак да нема систематске потребе за променом референтних вредности које је произвео модел.

5.2.4 Евалуација добијених лексикона на синтетичком евалуационом скупу

Након провере самог синтетичког скупа за евалуацију, представљеног у одељку 5.3.2, добијени лексикони сентимента упоређени су са његовим вредностима. Ручна провера осам анотатора показала је да се скуп може користити без измене референтних вредности модела, па је поређење изведено над изворним вредностима синтетичког скупа. Сваки од лексикона кандидата садржао је вредности за 25.320 синсета Српског ворднета, али су метрике израчунате само над 75 синсета који припадају синтетичком евалуационом скупу. На тај начин је омогућено да се целокупни лексикони S0-S7 упореде у истом контролисаном пресеку, без проблема непокривености, јер су сви идентификатори из скупа за евалуацију били присутни у свим кандидатским датотекама. Резултати поређења приказани су у Табели 5.6.

Табела 5.6 Поређење лексикона кандидата сентимента са синтетичким евалуационим скупом из одељка 5.2.3

Ранг	Лексикон	Породица модела	Средња апсолутна грешка	Просечно одступање са предзнаком	Тачно поклапање	Одступање до 0,25	Спирманова корелација	Квадратно пондерисана мера
1	S5	BERTić	0,124	-0,035	0,720	0,800	0,444	0,430
2	S7	Jerteh-355	0,127	-0,042	0,727	0,793	0,407	0,393
3	S6	GPT2-Ogao	0,133	-0,095	0,733	0,787	0,314	0,241
4	S0	SWN30 пресликавање	0,140	-0,140	0,760	0,800	није дефинисана	0,000
5	S4	компактни трансформер	0,150	-0,058	0,660	0,773	0,312	0,258
6	S3	RNN	0,152	-0,095	0,693	0,793	0,200	0,055

7	S2	AdaBoost + Бернулијев наивни Бајесов класификатор	0,152	-0,095	0,713	0,787	0,158	0,120
8	S1	SVM + Бернулијев наивни Бајесов класификатор	0,197	-0,015	0,493	0,693	0,170	0,117

У Табели 5.6 средња вредност апсолутне грешке означава просечно апсолутно одступање лексикона од вредности из синтетичког скупа за евалуацију. Просечно одступање са предзнаком (енгл. *mean signed deviation*) показује смер систематског одступања: негативна вредност значи да лексикон у просеку потцењује поларитет у односу на евалуациони скуп. Тачно поклапање и одступање до 0,25 рачунају се након заокруживања на скалу 0, 0,25, 0,5, 0,75, 1. Спирманова корелација показује да ли лексикон и евалуациони скуп слично рангирају синсете по јачини поларитета. Квадратно пондерисана к мера процењује сагласност на степенованој скали: одступање за један корак кажњава се мање него одступања за више корака.

Резултати показују да најмању средњу апсолутну грешку у односу на синтетички скуп за евалуацију има лексикон S5, добијен на основу модела BERTić. Његова комбинована вредност средње апсолутне грешке износи 0,124, уз Спирманову корелацију 0,444 и квадратно пондерисану к меру 0,4305. Веома близак резултат постиже S7, заснован на моделу Jerteh-355, са средњом апсолутном грешком 0,127, Спирмановом корелацијом 0,407 и квадратно пондерисаном к мером 0,3933. Трећи је S6, заснован на моделу GPT2-Orao, са средњом апсолутном грешком 0,1331.

Полазни лексикон S0, добијен директним пресликавањем из SentiWordNet-a, на први поглед постиже релативно високу вредност тачног поклапања (0,76) и исти резултат као S5 на мери одступања до 0,25 (0,8). Овај резултат, међутим, треба тумачити у светлу структуре синтетичког скупа за евалуацију. Сви синсети у овом скупу изабрани су из случајева који су у полазном SWN пресликавању имали вредност (0,0). Зато S0 за свих 75 синсета задржава нулте вредности и поклапа се са оним делом скупа за евалуацију који је и после процене модела остао објективан. Тај резултат је методолошки очекиван и показује да лексички сигнал из *Senti-pol-sr* није сам по себи доказ поларитета синсета. Ипак, S0 не разликује синсете који су и у скупу за евалуацију означени као неутрални од оних којима је у том скупу додељена ненулта вредност позитивне или негативне компоненте. Спирманова корелација за S0 није дефинисана, јер су вредности константне, а квадратно пондерисана к мера једнака је нули. То показује да S0 не успоставља корисно ординално (које се односи на поредак) рангирање (енгл. *ordinal ranking*) синсета према степену поларитета.

У том смислу, тачно поклапање није довољно за избор најбољег лексикона, јер је осетљиво на велики број нултих вредности у синтетичком евалуационом скупу. За ову евалуацију информативније су средња вредност апсолутне грешке, Спирманова корелација и квадратно пондерисана к мера, јер истовремено показују величину одступања и способност лексикона да очува степеновану структуру поларитета. Према тим мерама, најповољнији резултат има S5, док S7 представља врло блиску алтернативу.

У односу на S0, лексикон S5 смањује средњу апсолутну грешку за приближно 11,2%, док S7 смањује грешку за приближно 9,5%. Ово побољшање није велико у апсолутној вредности, али је методолошки значајно јер се јавља управо на подскупу синсета код којих је полазно

пресликавање било неутрално, а лексички сигнал из *Senti-pol-sr* указивао да је код одређених синсета поларитет сентимента ипак постоји. Другим речима, побољшање се не односи на лако препознатљиве снажно поларизоване случајеве, већ на оне синсете код којих је потребно разликовати објективно значење од значења са умереним или јачим поларитетом.

Сви лексикони, укључујући и најбоље рангиране, имају негативну вредност просечног потписаног одступања. То значи да у просеку додељују ниже вредности поларитета од вредности у синтетичком евалуационом скупу. Највеће негативно одступање забележено је код S0 и износи $-0,14$, што је очекивано јер су све вредности S0 у овом подскупу нулте. Код S5 и S7 просечно потписано одступање ближе је нули: и ови лексикони у просеку додељују ниже вредности од референтних, али је разлика мања.

Ови резултати допуњују ручну проверу синсета са највишим вредностима поларитета. Док је ручна провера у одељку 5.2.3 била усмерена на најизраженије позитивне и негативне случајеве, синтетички скуп за евалуацију из одељка 0 омогућава поређење целе скале вредности на контролисаном узорку. Због тога се разлике између резултата у 5.2.2 и 5.2.4 не тумаче као противречност, већ као последица различитих критеријума: један поступак мери поузданост најјачих вредности поларитета, а други степеновано поклапање са синтетичким евалуационим скупом.

Накнадна екстерна евалуација у анализи сентимента реченица, приказана у одељку 5.11, мења овај закључак, већ га поставља у шири оквир. На синтетичком евалуационом скупу најбољи је S5, уз врло близак S7, док у низводном задатку одређивања сентимента реченице најјачу потврду има S7. Зато резултате из овог одељка не треба механички прихватати као коначан ранг за све употребе ресурса: овај одељак вреднује усклађеност са вредностима синтетичког евалуационог скупа, док одељак 5.11 вреднује употребљивост слоја обележеног сентиментом у анализи реченица.

5.3 Изградња почетног инвентара значења WSD-sr-sens-inv-v1

Српски инвентар значења **WSD-sr-sens-inv** развијен је на основу ресурса Српски ворднет, који представља његову основну лексичко-семантичку основу. Међутим, пошто Српски ворднет не обезбеђује потпуну покривеност свих лексема и значења потребних за обележавање значења речи у ширем корпусу, било је неопходно његово систематско проширење вођено текстом који је одабран за обраду. Зато је изградња инвентара била непосредно повезана са препознавањем и укључивањем синсета који недостају у Српском ворднету (Petalinkar et al. 2026).

Као полазиште за утврђивање недостајућих синсета послужио је енглески инвентар значења ELEXIS-WSD-en, заснован на Принстонском ворднету. Од укупно 16.106 јединица у овом инвентару, поређењем дефиниција поравнато је 13.703 синсета из Принстонског ворднета. Иако је ELEXIS-WSD-en заснован на Принстонском ворднету, ELEXIS-WSD-en не садржи идентификаторе типа ILL у радном скупу је тај податак додат ради даљег упаривања са Српским ворднетом. У тренутку поравнања Српски ворднет је садржао 25.322 синсета и применом међујезичког индекса успешно је упарено 5.997 синсета између енглеског инвентара значења (приширеног са мапираним ILL) и Српског ворднета.

Преостали неупарени синсети затим су додатно испитани у односу на реченичне анотације у ресурсу ELEXIS-WSD-en. На тај начин издвојен је ужи, приоритетни скуп од 2.130 синсета за које је процењено да су највероватније неопходни за анотацију значења у корпусу. За овај скуп су синоними и дефиниције синсета из Принстонског ворднета аутоматски преведени на српски

језик применом **Google Translation API-ја**, модела **Gemini 2.5** и **GPT-3.5**, а затим је за сваку реченицу од превода које су понудили модели биран најбољи, преводи су ручно кориговани да би били у духу српског језика, да би по значењу одговорили оригиналу и да би били термилошки уједињени.

У следећој фази из скупа од 2.024 српске реченице издвојене су све пунозначне врсте речи, односно заједничке именице, глаголи, придеви и прилози. Оне лексеме које још увек нису биле заступљене у Српском ворднету превођене су на енглески језик како би се у Принстонском ворднету пронашли одговарајући кандидати. Овим поступком добијено је 1.795 додатних кандидата за синсете, који су затим аутоматски преведени и ручно дорађени. Поред тога, уведено је још 355 синсета који представљају хиперониме новододатих синсета, како би се очувала хијерархијска повезаност лексичке мреже. Укупно је у овој фази додато 2.150 синсета. У инвентар су унете само оне јединице које су, после поравнања, превођења и ручне дораде или лексикографске провере, оцењене као довољно поуздане за даљу употребу у аутоматској анотацији. После ових проширења Српски ворднет је обухватао 29.629 синсета, и то 21.948 именица, 3.695 глагола, 3.548 придева и 438 прилога. За потребе овог истраживања у инвентар су укључена само она значења која су релевантна за речи забележене у корпусу.

За преостале пунозначне речи којима ни после ових корака није било могуће придружити одговарајуће значење из инвентара, дефиниције су, кад год је то било могуће, преузимане из описних речника српског језика⁴⁷. То је најчешће био случај код придева и прилога који немају непосредан енглески еквивалент или нису заступљени у Принстонском ворднету, као што су, на пример, *дувански* „који се односи на дуван“ и *зими* „за време зиме, током зиме“. Посебну групу чиниле су именице женског рода изведене од именица мушког рода, на пример *новинарка* „женска особа новинар“, као и деминутиви и аугментативи, на пример *јајашице* „деминутив од јаје“. Поред тога, уочени су и појмови који уопште нису заступљени у Принстонском ворднету, као што је *дума* „скупштина, парламент у Русији“.

За лексеме за које ни после наведених поступака нису постојале одговарајуће дефиниције у српском инвентару значења, нове одреднице су обликоване применом техника инжењерства промптова уз употребу великих језичких модела, пре свега **Gemini (gemini-2.0-flash-exp)** и **OpenAI GPT-4.1**. Тако добијене дефиниције затим су ручно редиговане. Ова група јединица обухватала је пре свега термилошке јединице, лексеме новијег порекла и вишечлане изразе који нису били заступљени у енглеском Принстонском ворднету. Примери таквих јединица су *хадејк* „најранији геолошки еон у историји Земље, који је претходио архаику“ и *мастер-клас* „лекција или курс који води врхунски стручњак у одређеној области“.

Поред једночланих лексичких јединица, пажња је посвећена и вишечланим јединицама. Њихово укључивање у инвентар показало се као посебно важно за оне случајеве у којима значење није могуће поуздано или уопште извести из појединачних саставних делова изрази. У овој фази идентификовање и описивање таквих јединица у великој мери се и даље ослањало на ручну експертску проверу и корекцију, уз подршку различитих спољних лексичких и термилошких извора⁴⁸.

⁴⁷ 1) 209 дефиниција из Речника савременог српског језика, пројекат у развоју, под руководством проф. раде Сијовић развија Друштво за језичке ресурсе и технологије; 2) 4 дефиниције из Речника српског језика, Матица српска (2011), Милица Вујанић и други;

⁴⁸ Речник савременог српског језика, пројекат у развоју који води проф. др Рада Стијовић; Фразеолошки речник српског језика из 2012. године Ђорђа Оташевића.

Тако је оформљена прва верзија инвентара, означена као **WSD-sr-sens-inv-v1**. Њена основна намена била је да обезбеди почетни, оперативни инвентар значења за аутоматску анотацију корпуса, али и да послужи као полазна основа за даље допуне и исправке током ручне евалуације. Ова фаза карактеристична је за почетну изградњу инвентара, док се касније допуне и исправке у већој мери ослањају на ручно проверене грешке и циљану редакцију инвентара значења. Већ у овој фази било је јасно да је инвентар потребно посматрати као развојни ресурс, чија ће се структура и садржај постепено прецизирати у складу са резултатима анотације и анализом уочених недостатака.

Приликом дистрибуције и публиковања ресурса, инвентар значења представљен је у табеларном облику, а затим је аутоматски конвертован у базу знања засновану на повезаним подацима (*Linked Data*) у RDF/Turtle (TTL) формату. У тако формираној структури, на врху хијерархије налазе се класе које одговарају врстама речи, док су у оквиру сваке класе организоване појединачне леме. Свако значење леме представљено је као засебна инстанца (примерак) повезана са одговарајућом лемом и обogaћена идентификатором синсета, дефиницијом и другим релевантним метаподацима.

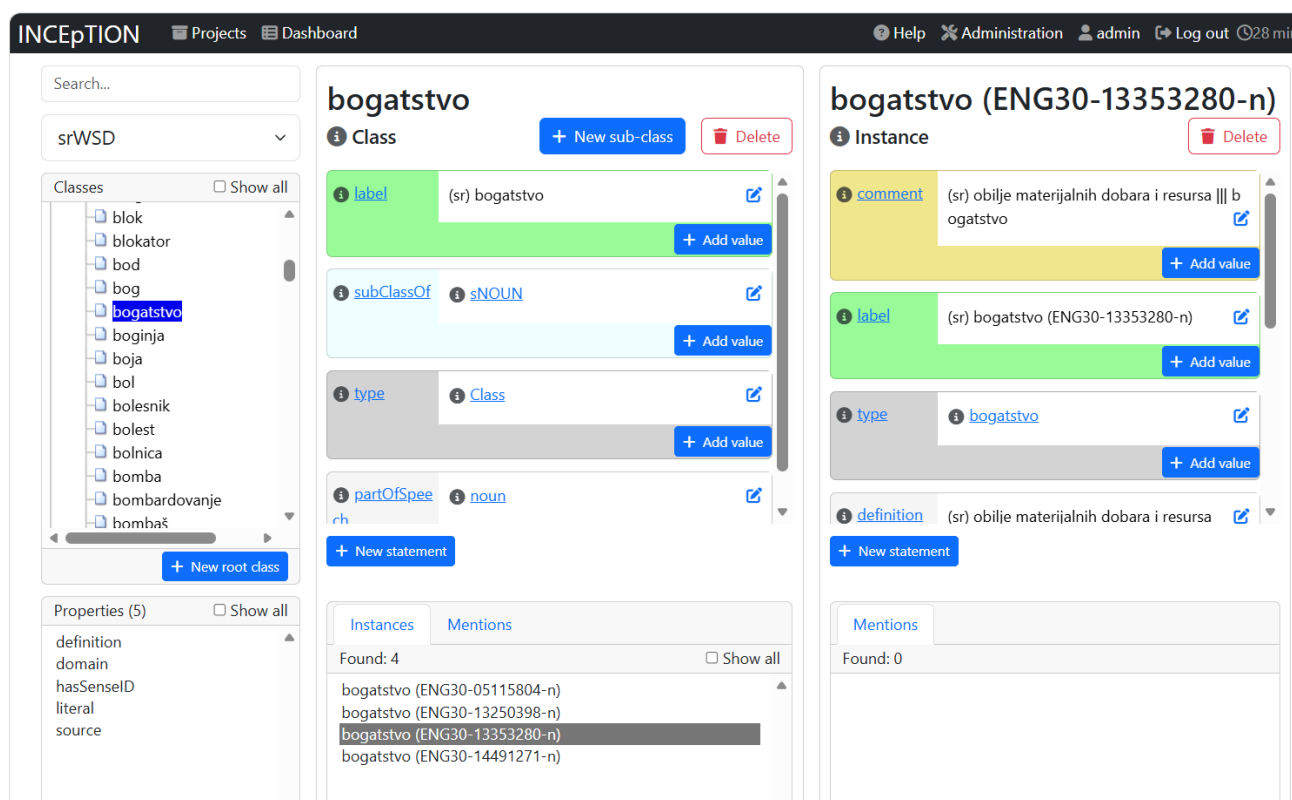
На основу тако генерисане TTL датотеке формирана је база знања која је увезена у алат **INCEpTION** (Klie et al. 2018), који омогућава визуелни преглед, претраживање и уређивање хијерархијски организованог инвентара значења. На Слици 5.2 приказан је пример представљања леме *богатство*, њених значења и припадајућих својстава у окружењу INCEpTION. Оваква организација ресурса омогућава једноставније управљање инвентаром значења, проверу конзистентности података и његову даљу употребу у задацима разрешења вишезначности речи и семантичке анотације.

Слика 5.2 приказује интерфејс алата **INCEpTION** који је прилагођен за уређивање и преглед инвентара значења заснованог на Српском ворднету. Приказан је пример лексеме „**богатство**“ и њених различитих значења (синсета).

Интерфејс је подељен у три главна дела:

- **Леви панел** садржи хијерархијски приказ класа, организованих по лемама. Изабрана је класа *богатство*, а испод су наведене и друге лексичке јединице. У доњем делу налази се листа доступних својстава (нпр. *definition*, *domain*, *hasSenseID*, *literal*, *source*).
- **Средишњи панел** приказује метаподатке класе *богатство*. Наведена је ознака леме (*label*), информација да припада категорији именица (*subClassOf sNOUN*), као и друга структурна својства. У доњем делу налази се листа свих значења (инстанци) повезаних са овом лемом. За лему *богатство* регистрована су четири синсета.
- **Десни панел** приказује детаље изабраног значења *богатство* (*ENG30-13353280-n*). Наведена је дефиниција „обиље материјалних добара и ресурса“, идентификатор синсета и припадност класи *богатство*. Овај панел омогућава уређивање и допуњавање појединачних значења.

Слика илуструје на који начин је организован инвентар значења током истраживања: свакој лемџ је придружен скуп синсета са јединственим идентификаторима и одговарајућим дефиницијама. Оваква организација омогућава систематско управљање значењима речи и њихову каснију примену у задацима разрешења вишезначности речи (WSD).



Слика 5.2 Интерфејс алата INCEpTION

5.4 Унапређење инвентара и његова улога у анотацији

После првог круга ручне корекције, ажурирана верзија инвентара објављена је као **WSD-sr-sens-inv-v2** и потом употребљена у новом циклусу аутоматске анотације. Ова, друга, ревидирана верзија обухватила је значења која су недостајала у корпусу и која су у претходној фази била обележавана привременом ознаком **NEW_SENSE** која је указивала да модел није могао из постојећег инвентара (**v1**) да понуди ниједно значење. Поред тога, у другу, ревидирану верзију инвентара укључена су и нова значења неопходна за разрешавање погрешно додељених значења уочених у првом развојно-аналитичком подскупу од 120 ручно проверених реченица, до чега је дошло јер у инвентару није постојало одговарајуће значење.

Посебан значај у овом поступку имала су структурирана образложења добијена током избора значења заснованог на великим језичким моделима. Она су олакшала ручну редакцију, јер су јасно показивала зашто је одређено значење изабрано, али су истовремено указивала и на случајеве у којима постојеће дефиниције или границе између значења нису биле довољно прецизне. Увођењем недостајућих значења друга, ревидирана верзија инвентара омогућила је прецизније аутоматско придруживање синсета у наредној фази анотације.

Поред додавања нових значења, исправљене су и постојеће одреднице. Укупно је побољшано 213 дефиниција. Осим исправљања типографских и граматичких пропуста, посебна пажња посвећена је дефиницијама које су биле недовољно информативне или недовољно прецизне. Тако је, на пример, дефиниција лексеме *девојчица* „млада женска особа“ замењена прецизнијом формулацијом „млада женска особа која још није девојка“, док је дефиниција глагола *учествовати* „бити укључен у нешто“ замењена формулацијом „узети, узимати учешћа у некој заједничкој акцији, суделовати“.

Поређење прве верзије инвентара и друге, ревидиране верзије, приказано у Табели 5.7, показује доследно, али циљано проширење инвентара значења. Највећи апсолутни пораст забележен је у броју различитих јединица у репозиторијуму и броју различитих значења у њему, што одражава систематско додавање значења чији недостатак је уочен током ручне евалуације и анализе грешака. Насупрот томе, број различитих лема и комбинација **UPOS+lemma** повећан је само незнатно, што указује на то да је тежиште редакције било на проширивању семантичке покривености већ постојећих лексичких јединица, а не на увођењу већег броја нових лема. Такав налаз у складу је и са начином на који је прошириван Српски ворднет. Слично томе, готово непромењен број једночланих и вишечланих израза показује да је редакција била усмерена пре свега на повећање семантичке прецизности и диференцираности инвентара, а не на промену његовог укупног лексичког састава.

Табела 5.7 Поређење броја одредница, значења, лема и других показатеља у верзијама *WSD-sr-sens-inv-v1* и *WSD-sr-sens-inv-v2*

Показатељ	v1	v2	Апсолутна разлика	%
Јединице у репозиторијуму	12 747	13 227	480	3,8
Значења	9 929	10 284	355	3,6
Леме	6 108	6 122	14	0,2
UPOS+lemma	6 122	6 136	14	0,2
– једночлане јединице	5 085	5 096	11	0,2
– вишечлане јединице	1 037	1 040	3	0,3

Детаљнији увид у промене код једночланих одредница дат је у Табели 5.8, где су подаци разложени по врстама речи. Највећи апсолутни пораст забележен је код именица, које су добиле 211 нових значења у репозиторијуму и седам нових лема. То потврђује њихову централну улогу у анотацији значења и указује на потребу да се управо у номиналном делу инвентара побољша покривеност уочена током ручне евалуације. Код придева и прилога раст је минималан, док су глаголи остали готово непромењени, што упућује на то да је глаголска покривеност већ у првој верзији инвентара била релативно стабилна. У целини посматрано, ови подаци потврђују да је прелаз са прве верзије инвентара на другу, ревидирану верзију, био пре свега усмерен на допуну и прецизирање инвентара значења за постојеће леме, а не на увођење већег броја нових једночланих лексичких јединица.

Како би проценили вишезначност речи у инвентару, урађена је анализа број значења за појединачне речи, што је потом груписано према броју компоненти и према врсти речи. У последњој колони Табеле 5.8 дат је и однос броја значења и броја лема у коначној верзији. Највећа вредност забележена је код глагола (3,8), а најмања код придева (1,8), док просек износи 2,3 значења по леми. Приказани подаци се слажу са опсервацијама анотатора да су код глагола били највећи изазови при селекцији одговарајућег значења.

Табела 5.8 Разлике код једночланих одредница по категоријама *UPOS* у верзијама *WSD-sr-sens-inv-v1* и *WSD-sr-sens-inv-v2*

UPOS	јединице значења v1	Јединице значења v2	Разлика	Леме v1	Леме v2	Разлика	Однос значења и лема у v2
NOUN	5 279	5 490	211	2 432	2 439	7	2,3

VERB	3 179	3 281	102	867	868	1	3,8
ADJ	2 565	2 686	121	1 527	1 529	2	1,8
ADV	480	512	32	259	260	1	2,0
Укупно	11 503	11 969	466	5 085	5 096	11	2,3

Сличан увид у промене код вишечланих одредница дат је у Табели 5.9. Пораст је и у овом случају био очекивано мали и када су у питању јединице значења и леме јер су вишечлане одреднице ретко кад вишезначне и већ првом верзијом инвентара су потребна значења била обухваћена: забележено је свега 14 нових значења и три нове леме. Највећи број значења по лема јавља се код глаголских вишечланих јединица (1,5), док су код именица и прилога вредности нешто ниже (1,1), а код вишечланих придева није забележен пораст, што је и очекивано с обзиром на њихову релативно малу заступљеност у ворднету. Ови резултати додатно потврђују да је редакција ресурса била усмерена на допуну и прецизирање већ постојећих делова инвентара, а не на значајније обогаћивање скупа вишечланих лексичких јединица.

Табела 5.9 Разлике код вишечланих одредница по категоријама UPOS у верзијама WSD-sr-sens-inv-v1 и WSD-sr-sens-inv-v2

UPOS	јединица значења v1	Јединица значења v2	Разлика	Леме v1	Леме v2	Разлика	Однос значења и лема у v2
NOUN	687	696	9	648	651	3	1,1
VERB	504	507	3	337	337	0	1,5
ADJ	5	5	0	5	5	0	1,0
ADV	48	50	2	47	47	0	1,1
Укупно	1 244	1 258	14	1 037	1 040	3	1,2

На основу приказаних података може се закључити да је прелаз са прве верзије инвентара на другу, ревидирану верзију, био усмерен пре свега на семантичко употпуњавање и прецизирање постојећег инвентара. Такве измене имале су непосредан значај за наредни круг аутоматске анотације, јер су омогућиле бољу покривеност недостајућих значења и смањиле број случајева у којима је систем био принуђен да прибегне привременој ознаци **NEW_SENSE** јер није могао у инвентару да пронађе одговарајуће значење. На тај начин друга верзија инвентара није представљала коначан ресурс, већ стабилизовану радну основу за евалуацију текуће итерације и за даље, контролисано проширивање инвентара значења.

5.5 Опис златног скупа и анализа инвентара значења

Претходни одељак приказује како је почетни инвентар значења проширен и ревидиран на основу аутоматске анотације, ручне провере и анализе случајева у којима постојећи инвентар није обезбеђивао одговарајуће значење. У овом одељку фокус је на ручно провереном евалуационом материјалу који је коришћен за процену таквог инвентара и за анализу тежине задатка разрешења вишезначности речи.

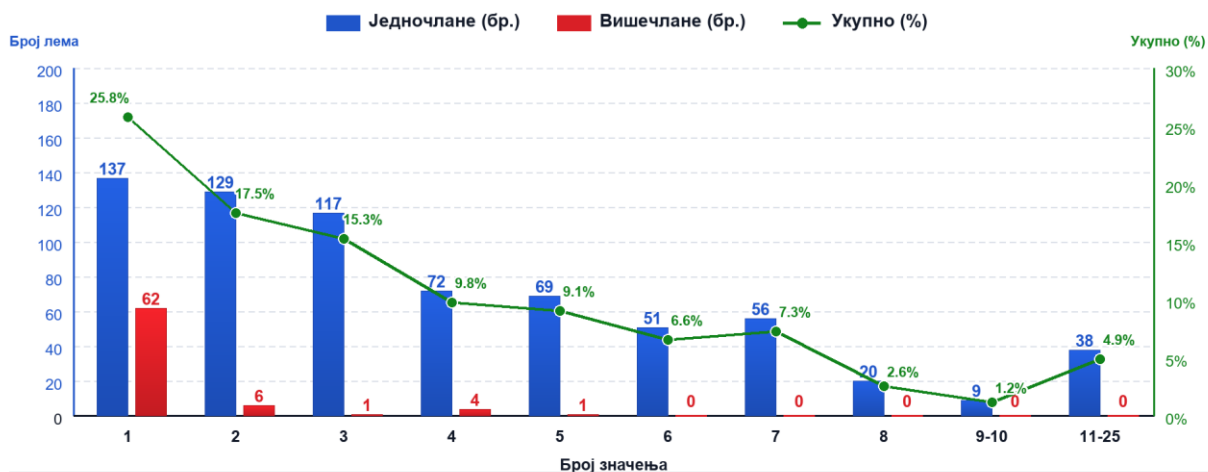
Као резултат овог процеса формиран је златни скуп ELEXIS-WSD-sr-gold. Он обухвата укупно 420 ручно проверених реченица, али тај материјал нема исту методолошку улогу у свим фазама поступка. Првих 120 реченица чини развојно-аналитички скуп: коришћен је за проверу

почетних резултата, анализу грешака и редакцију инвентара. Наредних 300 реченица чини издвојени тест скуп за процену текуће верзије поступка, јер није коришћен за подешавање модела, упита или инвентара након прве итерације.

Укупно посматрано, златни скуп садржи 2.693 аотирана сегмента текста којима је додељено значење из инвентара. Од тога је 2.430 једночланих, а 263 вишечланих јединица. Развојно-аналитички скуп садржи 772 аотирана сегмента текста, од чега 698 једночланих и 74 вишечлана. Издвојени тест скуп садржи 1.921 аотирани сегмент текста, од чега 1.734 једночланих и 187 вишечланих. Посматрано по различитим лексичким јединицама, скуп обухвата 583 различите једночлане јединице и 65 различитих вишечланих јединица, што указује на релативно широку лексичку покривеност у оквиру ручно провереног материјала.

Анализа учесталости показује да су најзаступљенија значења у оквиру једночланих јединица „држава, земља“ (једанаест појављивања) при чему „држава“ има три различита значења, затим „млад, незрео“ (осам појављивања), као и глагол „моћи“ у значењу „имати могућност“ (седам појава). Посебну пажњу захтева управо глагол „моћи“, јер у енглеском језику често одговара модалном глаголу *can*, који као помоћни глагол није представљен у оквиру стандардних ресурса типа ворднет. У Српском ворднету је за овај глагол било доступно само једно значење повезано са енглеским синсетом, док је у процесу редакције додато још пет значења. Овај пример илуструје ограничења постојећих лексичких ресурса и потребу за њиховим прилагођавањем специфичностима српског језика. Када су у питању вишечлане јединице, најчешће су „јавити се“ у значењу „појавити се“ са три појављивања у корпусу, а затим „фудбалски тим“ и „размена информација“, са по два појављивања. Иако су бројчано мање заступљене од једночланих јединица, ове јединице имају важну улогу у анализи, јер често носе специфична значења која се не могу директно извести из појединачних компоненти.

Свакој аотираној јединици придружен је скуп могућих значења, одређен на основу комбинације леме и врсте речи. Због тога број кандидата за значење представља важан показатељ тежине задатка. Расподела броја понуђених значења у развојно-аналитичком скупу приказана је на Слици 5.3, док је одговарајућа расподела за издвојени тест скуп приказана на Слици 5.4. Одвојено приказивање ова два дела златног скупа је важно јер они имају различиту улогу у итеративном поступку.

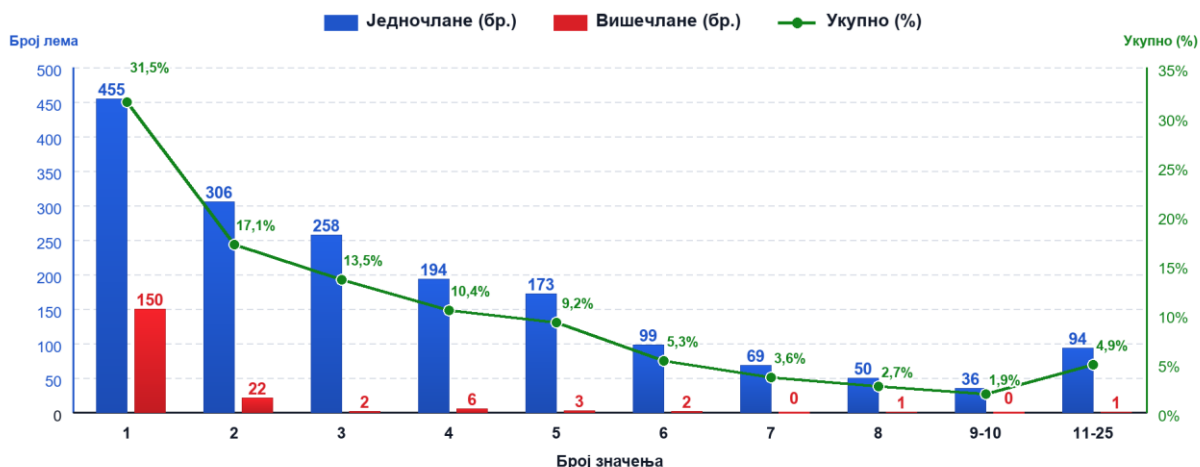


Број значења	1	2	3	4	5	6	7	8	9-10	11-25	Укупно
Једночлане (бр.)	137	129	117	72	69	51	56	20	9	38	698
Једночлане (%)	19.6	18.5	16.8	10.3	9.9	7.3	8.0	2.9	1.3	5.4	100.0
Вишечлане (бр.)	62	6	1	4	1	0	0	0	0	0	74
Вишечлане (%)	83.8	8.1	1.4	5.4	1.4	0.0	0.0	0.0	0.0	0.0	100.0
Укупно (бр.)	199	135	118	76	70	51	56	20	9	38	772
Укупно (%)	25.8	17.5	15.3	9.8	9.1	6.6	7.3	2.6	1.2	4.9	100.0

Напомена: Проценти су израчунати у оквиру сваке групе посебно (једночлане, вишечлане) и у односу на укупан број лема (772).

Слика 5.3 Анотирани сегменти текста према броју понуђених могућих значења у развојном скупу (120 реченица).

Као што се види на Слици 5.3, у развојно-аналитичком скупу највећи број анотираних сегмената текста припада групама са бројем понуђених значења од један до пет, док су сегменти текста са више од десет значења кандидата знатно ређи. Та расподела показује да први ручно проверени скуп садржи довољно примера за анализу уобичајених случајева разрешења вишезначности, али и мањи број изразито полисемичних јединица погодних за анализу грешака и редакцију инвентара. Иста анализа примењена је и на издвојени тест скуп од 300 реченица.



Број значења	1	2	3	4	5	6	7	8	9-10	11-25	Укупно
Једночлане (бр.)	455	306	258	194	173	99	69	50	36	94	1.734
Једночлане (%)	26,2	17,6	14,9	11,2	10,0	5,7	4,0	2,9	2,1	5,4	100,0
Вишечлане (бр.)	150	22	2	6	3	2	0	1	0	1	187
Вишечлане (%)	80,2	11,8	1,1	3,2	1,6	1,1	0,0	0,5	0,0	0,5	100,0
Укупно (бр.)	605	328	260	200	176	101	69	51	36	95	1.921
Укупно (%)	31,5	17,1	13,5	10,4	9,2	5,3	3,6	2,7	1,9	4,9	100,0

Напомена: Проценти су израчунати у оквиру сваке групе посебно (једночлане, вишечлане) и у односу на укупан број лема (1.921).

Слика 5.4 Анотирани сегменти текста према броју понуђених могућих значења у издвојеном тест скупу (300 реченица).

Подаци на скилама 5.3 и 5.4 показују да већина анотираних јединица има релативно ограничен број кандидатских значења. У развојно-аналитичком скупу, 199 од 772 сегмента текста има само једно понуђено значење, док у издвојеном тест скупу ту групу чини 605 од 1.921 сегмента текста. Уочава се да јединице које имају највише пет кандидатских значењачине највећи део оба скупа. То значи да задатак у већини случајева не подразумева избор међу великим бројем кандидата, али ипак захтева прецизно повезивање контекста са одговарајућим значењем.

С друге стране, мањи број јединица показује изражену полисемију⁴⁹ што се одражава и бројем значења забележених у инвентару, а управо тај део скупа има посебну методолошку вредност. Тако, на пример, глагол „дати“ има 25 могућих значења у оквиру инвентара, а глагол „добити“ 19 значења. Такви случајеви представљају највећи изазов за разрешење вишезначности речи, јер захтевају фину контекстуалну интерпретацију и осетљивост на разлике између блиских дефиниција.

Слика 5.5 приказује радно окружење INCErTION за анотацију корпуса у задатку разрешења вишезначности речи (WSD). У тексту су означени сви токени, при чему су корпусним речима (једночланим и вишечланим) придружене њихове леме и значење, при чему свака анотација садржи идентификатор значења, извор и додатне метаподатке. Именовани ентитети су такође анотирани, класификовани где је то било могуће повезани са сопољашњом базом знања, Посебно је илустрован механизам повезивања анотираних речи са одговарајућим описом из инвентара значења.

⁴⁹ Полисемија (вишезначност) јесте појава да једна лексема има више међусобно повезаних значења, која се остварују у различитим контекстима.

Посебно је упечатљив пример из реченице 8, у којој се јавља реч *програма* у контексту исказа „Резултат друге сезоне програма Фудбал за пријатељство био је избор лидера покрета“. За ову појаву лексеме систем је предложио више могућих значења (укупно седам), а као најприкладније изабрано је значење са идентификатором ENG30-05899087-n. Кликом на анотацију отвара се прозор у коме се поред детаљних података из инвентара, налази и објашњење избора, број доступних значења, и листа понуђених кандидата.

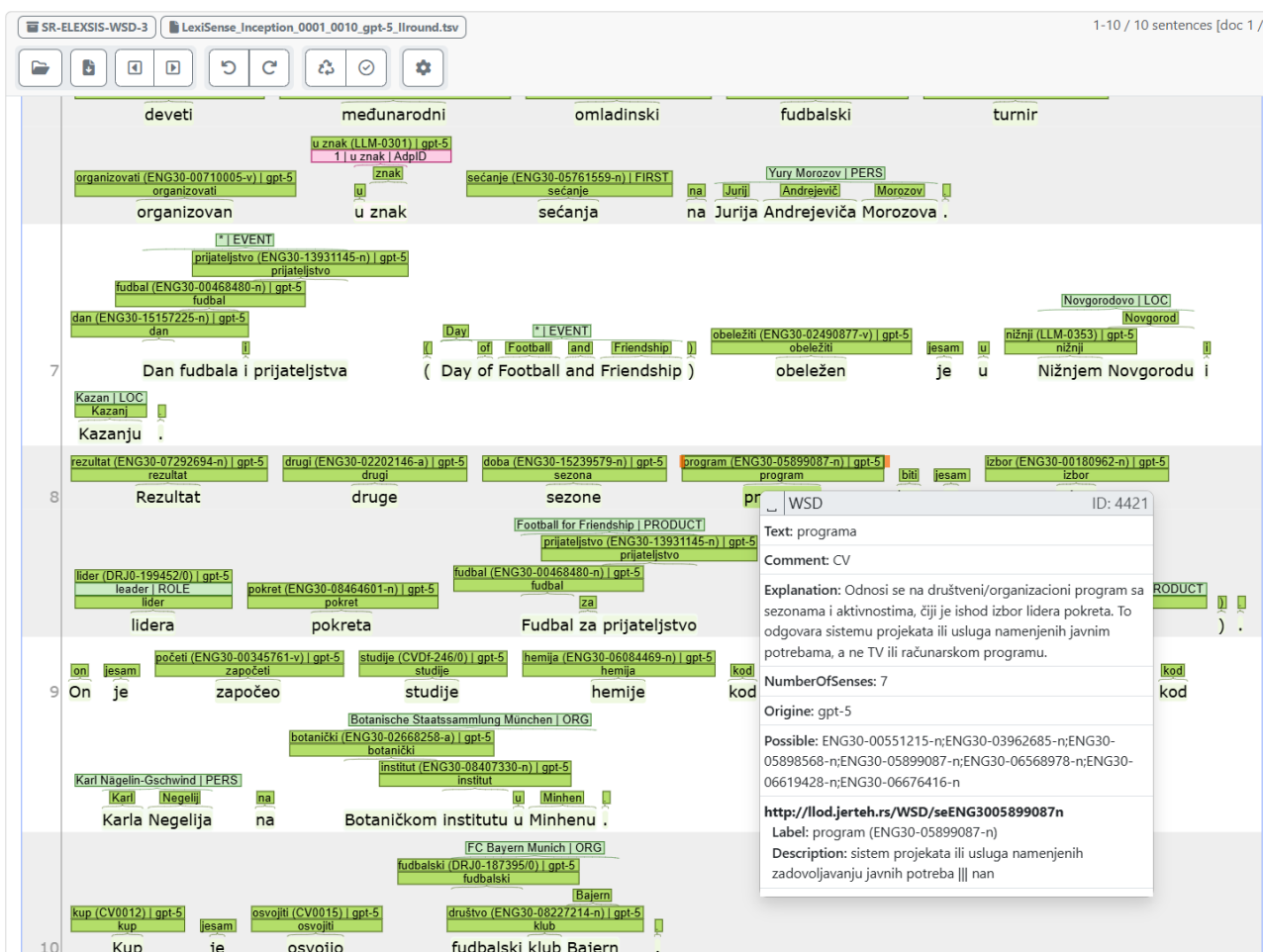
У пољу Explanation наведено је образложење да се реч *програм* у датом контексту односи на друштвени или организациони програм са активностима и сезонама чији је исход избор лидера покрета, а не на телевизијски или рачунарски програм. Овакво образложење представља важан елемент процеса анотације јер експлицитно повезује контекст реченице са семантичким својствима изабраног значења.

Инвентар за изабрано значење даје следећи опис:

„систем пројеката или услуга намењених задовољавању јавних потреба“.

Ова дефиниција је у складу са контекстом реченице, јер се назив *Фудбал за пријатељство* не односи на појединачан догађај нити на рачунарски програм, већ на организовану иницијативу која обухвата више активности, сезона и учесника. На основу тога може се закључити да је изабрано значење семантички оправдано и да показује како комбинација контекстуалних информација и знања садржаног у инвентару омогућава успешно идентификовање значења.

Овај пример истовремено илуструје једну од предности предложеног приступа: анотатор не добија само идентификатор значења већ и објашњење избора засновано на описима из инвентара и контексту реченице, што омогућава транспарентнију валидацију резултата и лакше уочавање потенцијалних грешака у процесу аутоматске анотације.



Слика 5.5 Радно окружење INCEpTION за анотацију корпуса

У целини посматрано, опис златног скупа показује две важне карактеристике евалуационог материјала. Прво, ручно проверени скуп је довољно разноврстан да обухвати и једночлане и вишечлане јединице, као и њихов различит степен вишезначности нивое полисемије, односно броја кандидата за значења која се нуде за одређену лему и врсту речи. Друго, издвајање развојно-аналитичког и тест дела омогућава да се резултати у наредном одељку не тумаче само као поређење модела, већ и као процена утицаја редакције инвентара значења на успешност аутоматске анотације значења.

5.6 Квантитативна евалуација перформанси

Квантитативна евалуација спроведена је у две итерације, које одговарају двема верзијама инвентара значења. У првој итерацији коришћен је инвентар WSD-sr-sens-inv-v1, док је у другој итерацији коришћена ревидирана верзија WSD-sr-sens-inv-v2 (5.3, 5.4) . На тај начин евалуација не мери само разлике између модела, већ и утицај проширења и корекције инвентара значења на успешност разрешења вишезначности речи.

Резултати су приказани одвојено за два ручно проверена скупа: развојно-аналитички и издвојени тест скуп који су представљени у претходном одељку. Исходи су сведени на три категорије: **тачно**, када је модел изабрао одговарајуће понуђено значење; **погрешно**, када је изабрао неодговарајуће понуђено значење; и **недедељено**, када није издвојено ниједно од

понуђених значења. У изворним резултатима ове категорије одговарају ознакама *matched*, *mismatched* и *unmatched*. У главним табелама приказане су процене тачности са 95% Вилсоновим интервалима поверења, док графички прикази илуструју расподелу тачних, погрешних и недодељених класификација.

Пре приказа резултата потребно је раздвојити улоге испитаних поступака. Ознаке у табелама не представљају увек засебне моделе: *base-lesk* је полазни поступак заснован на Лесковом алгоритму, *simple*, *tesla* и *multilingual-e5-large* су енкодерски модели за семантичко упаривање контекста и кандидатског значења, док су *Llama4*, *Mistral-Small-3.2*, *GPT* и *Gemini* генеративни модели коришћени у поступку разрешења вишезначности речи. Ознаке *Llama4-ab1*, *Llama4-ab2* и *Llama4-ab3* не означавају нове породице модела, већ аблационе варијанте промпта за исти *Llama4* модел. Ова подела је сажета у Табели 5.10.

Табела 5.10 Улога модела и варијанти поступка у евалуацији разрешења вишезначности речи

Ознака у табелама	Пуно име модела или поступка	Тип	Улога у евалуацији
<i>base-lesk</i>	Лесков алгоритам (Lesk, 1986)	полазни поступак заснован на знању	Полазни резултат заснован на преклапању контекста и глоса кандидатских значења; не представља модел заснован на неуронским мрежама.
<i>simple</i>	<i>sentence-transformers/all-MiniLM-L6-v2</i>	енкодерски модел реченица	Компактан модел за реченичне репрезентације; једноставни енкодерски полазни модел.
<i>tesla</i>	<i>te-sla/TeslaXLM</i>	вишејезични енкодер	Модел прилагођен српском и српскохрватском материјалу; провера користи од језички ближе репрезентације.
<i>multilingual-e5-large</i>	<i>intfloat/multilingual-e5-large</i>	вишејезични E5 енкодер	Модел за семантичко упаривање контекста и кандидатског значења.
<i>Llama4</i>	<i>meta-llama/Llama-4-Scout-17B-16E-Instruct</i> ; локално <i>llama4:latest</i>	генеративни модел	Основни <i>Llama</i> поступак са пуним промptom, опцијом <i>NEW_SENSE</i> и образложењем избора.
<i>Llama4-ab1</i>	исти <i>Llama4</i> модел, без захтева за образложењем	аблација упита	Провера утицаја образложења; излаз садржи само изабрани <i>sense_id</i> .
<i>Llama4-ab2</i>	исти <i>Llama4</i> модел, без опције <i>NEW_SENSE</i>	аблација упита	Провера утицаја могућности уздржавања; модел мора да изабере једно од понуђених значења.
<i>Llama4-ab3</i>	исти <i>Llama4</i> модел, са скраћеним промptom	аблација упита	Провера утицаја дужине промпта; задржани су <i>NEW_SENSE</i> и образложење.

Ознака у табелама	Пуно име модела или поступка	Тип	Улога у евалуацији
Mistral-Small-3.2	mistralai/Mistral-Small-3.2-24B-Instruct-2506; локално mistral-small3.2:latest	генеративни модел	Додатни отворени генеративни модел за поређење са Llama4 и затвореним API моделима.
GPT модели	OpenAI GPT модели коришћени преко API-ја	затворени генеративни модели	Поређење са комерцијалним моделима високих перформанси.
Gemini модели	Google Gemini модели	затворени генеративни модели	Поређење са Google генеративним моделима у истом задатку.

Као што показује Табела 5.10, резултати у наставку не представљају једноставно рангирање хомогене групе модела. У истом прегледу налазе се полазни поступак заснован на знању, енкодерски модели, генеративни модели и аблације упита. Због тога је при тумачењу резултата важно раздвојити поређење породица модела од провере стабилности појединачног промпта.

На развојно-аналитичком скупу од 120 реченица друга итерација доводи до побољшања код већине модела. Побољшање је видљивије код једночланих јединица, где је простор избора шири и где непотпуност инвентара непосредније утиче на исход. Резултати за једночлане јединице, које су и изазовнији проблем, приказани су у Табели 5.11.

Табела 5.11 Тачност модела на једночланим јединицама у развојно-аналитичком скупу од 120 реченица, где R1 и R2 означавају прву и другу итерацију евалуације

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
base-lesk	37,0% (33,5-40,6)	46,1% (42,5-49,8)	+9,1
simple	37,7% (34,2-41,3)	44,8% (41,2-48,6)	+7,1
tesla	35,8% (32,3-39,4)	48,7% (45,0-52,4)	+12,9
multilingual-e5-large	40,7% (37,1-44,4)	54,3% (50,6-58,0)	+13,6
Llama4	54,0% (50,3-57,7)	66,8% (63,2-70,2)	+12,8
Llama4-ab1	51,3% (47,6-55,0)	60,0% (56,3-63,6)	+8,7
Llama4-ab2	54,7% (51,0-58,4)	67,6% (64,1-71,0)	+12,9
Llama4-ab3	52,6% (48,9-56,3)	65,6% (62,0-69,0)	+13,0
Mistral-Small-3.2	57,2% (53,5-60,8)	74,1% (70,7-77,2)	+16,9
GPT-3.5	53,2% (49,4-56,8)	-	-

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
GPT-4.1	66,2% (62,6-69,6)	81,5% (78,5-84,2)	+15,3
GPT-5.1	-	79,9% (76,8-82,7)	-
Gemini-2.5 Flash	56,4% (52,7-60,1)	-	-
GeminiPro-2.5	60,0% (56,3-63,6)	78,2% (75,0-81,1)	+18,2

Табела 5.11 показује да је побољшање између две итерације присутно и код полазних и код јачих генеративних модела, али није истог обима. Код енкодерских модела добитак је умерен, док је код GPT-4.1 и GeminiPro-2.5 знатно израженији. То указује да редакција инвентара не делује изоловано, већ највише долази до изражаја када модел има довољну способност да искористи прецизније дефинисана кандидатска значења.

Код вишечланих јединица, приказаних у Табели 5.12, почетне вредности су више него код једночланих јединица. То је очекивано, јер је њихово значење често специфичније и немају већи број стварно конкурентних тумачења.

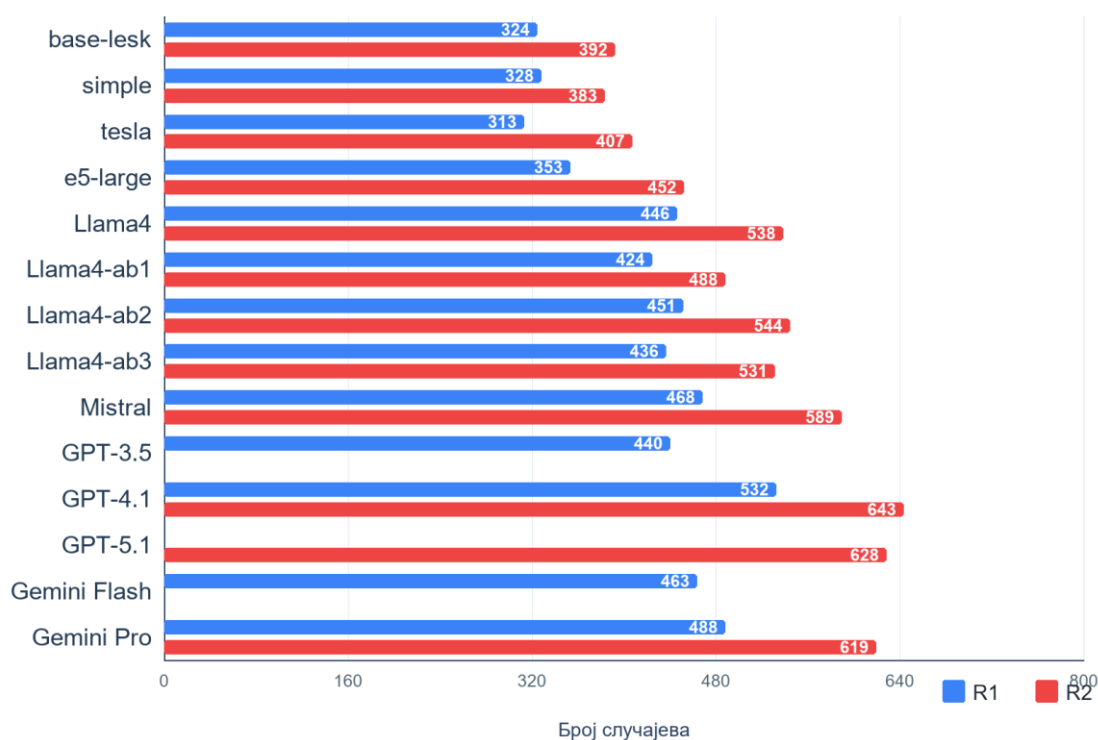
Табела 5.12 Тачност модела на вишечланим јединицама у развојно-аналитичком скупу од 120 реченица

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
base-lesk	89,2% (80,1-94,4)	94,6% (86,9-97,9)	+5,4
simple	87,8% (78,5-93,5)	94,6% (86,9-97,9)	+6,8
tesla	85,1% (75,3-91,5)	90,5% (81,7-95,3)	+5,4
multilingual-e5-large	93,2% (85,1-97,1)	98,6% (92,7-99,8)	+5,4
Llama4	93,2% (85,1-97,1)	97,3% (90,7-99,3)	+4,1
Llama4-ab1	89,2% (80,1-94,4)	93,2% (85,1-97,1)	+4,0
Llama4-ab2	93,2% (85,1-97,1)	97,3% (90,7-99,3)	+4,1
Llama4-ab3	93,2% (85,1-97,1)	98,6% (92,7-99,8)	+5,4
Mistral-Small-3.2	93,2% (85,1-97,1)	97,3% (90,7-99,3)	+4,1
GPT-3.5	93,2% (85,1-97,1)	-	-
GPT-4.1	94,6% (86,9-97,9)	100,0% (95,1-100,0)	+5,4
GPT-5.1	-	94,6% (86,9-97,9)	-

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
Gemini-2.5 Flash	93,2% (85,1-97,1)	-	-
GeminiPro-2.5	93,2% (85,1-97,1)	98,6% (92,7-99,8)	+5,4

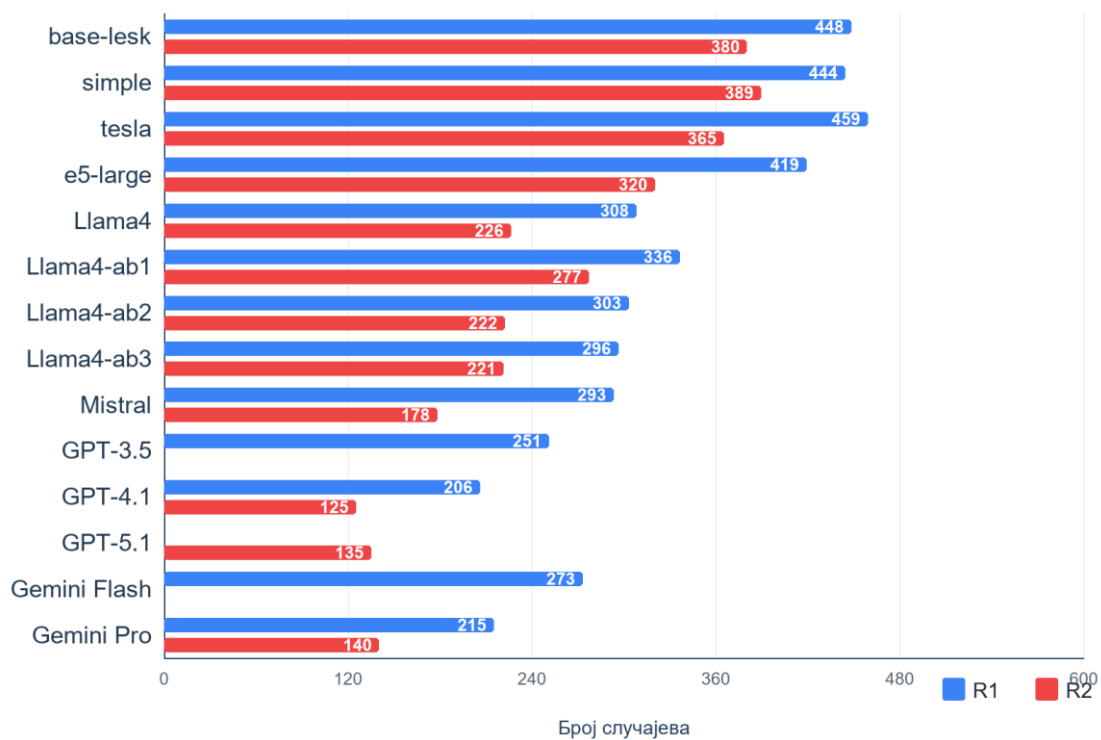
Табела 5.12 потврђује да вишечлане јединице нису главни извор разлике између модела: већина система постиже високе резултате, а интервали поверења су шири због мањег броја примера. Због тога је за развојно-аналитички скуп важнији налаз из Табеле 5.11, где се види јаснија разлика између полазних поступака, енкодера и генеративних модела.

Табеле 5.11 и 5.12 приказују тачност, али не показују како је грешка распоређена између погрешног избора и недодељеног значења. Због тога су на сликама 5.6, 5.7 и 5.8 приказани бројеви тачних, погрешних и недодељених исхода на развојно-аналитичком скупу. Слика 5.6 показује раст броја тачних исхода после редакције инвентара, Слика 5.7 показује смањење броја погрешних избора, док Слика 5.8 издваја недодељене исходе, који су посебно важни као сигнал непокривености или неусклађености инвентара.



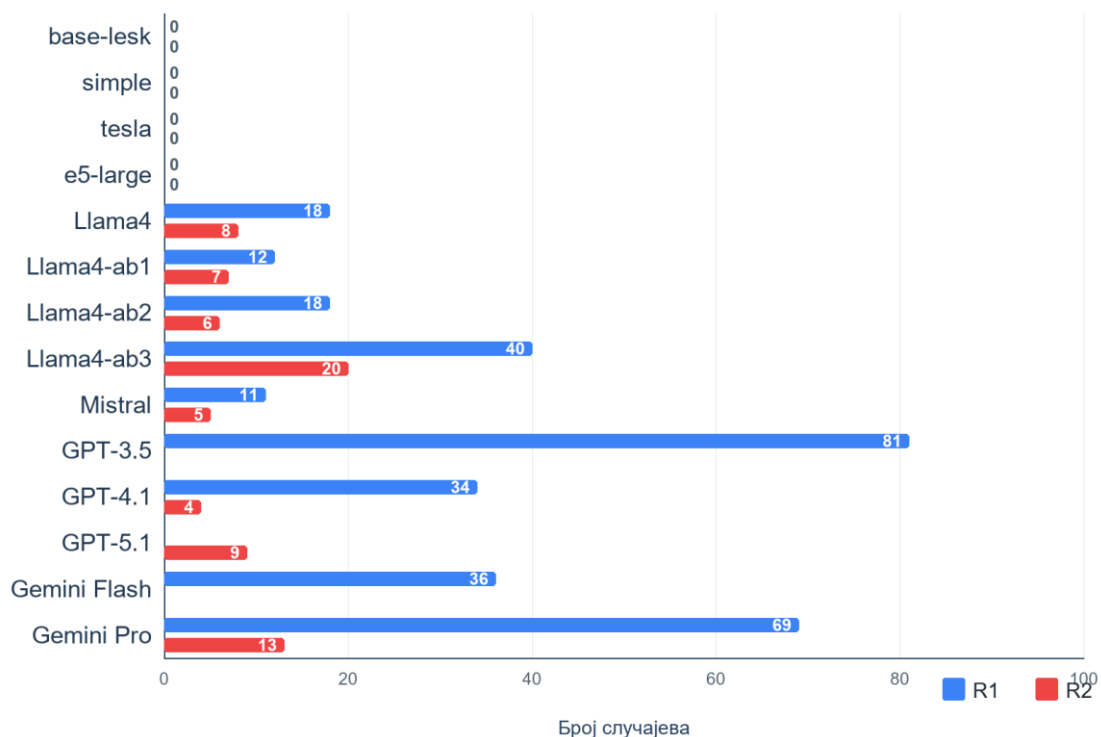
Слика 5.6 Број тачно разрешених случајева на развојно-аналитичком скупу од 120 реченица

На развојно-аналитичком скупу број тачно разрешених случајева расте у другој итерацији код свих поступака који су упоредиви у оба круга. Највећи број тачних исхода у R2 има GPT-4.1 (643), затим GPT-5.1 (628) и Gemini Pro (619), а пораст је код модела с највећим бројем тачних исхода GPT-4.1 у односу на R1 111 случајева. Овај приказ зато не показује само разлике између модела, већ и утицај редакције инвентара значења на скупу који је имао развојно-аналитичку улогу.



Слика 5.7 Број погрешно разрешених случајева на развојно-аналитичком скупу од 120 реченица.

Број погрешно разрешених случајева на истом скупу опада у R2, што је очекивано после допуне и прецизирања инвентара значења. Код GPT-4.1 број погрешних исхода смањује се са 206 на 125, а код Mistral-Small-3.2 са 293 на 178. Ипак, код полазних и једноставнијих поступака број погрешних избора остаје и даље висок, на пример 380 код base-lesk и 389 код simple у R2.



Слика 5.8 Број недодељених случајева на развојно-аналитичком скупу од 120 реченица.

Недодељени исходи су посебно важни јер указују на случајеве у којима модел није издвојио ниједно од понуђених значења. На развојно-аналитичком скупу њихов број у R2 нагло опада код већине генеративних модела: код GPT-4.1 са 34 на 4, код Gemini Pro са 69 на 13, а код Llama4 са 18 на 8. Такав пад је у складу са улогом друге итерације, у којој је инвентар циљано допуњаван управо на основу анализе почетних недостатака.

Графикони на сликама 5.6, 5.7 и 5.8 показују да се побољшање у другој итерацији не може посматрати само кроз проценат успешности, већ треба посматрати и структуру исхода. Код јачих генеративних модела расте број тачних избора, док се недодељени исходи смањују проширивањем инвентара и његовим бољим усклађивањем са корпусним примерима. Овај приказ је користан јер показује оно што у табелама тачности остаје прикривено: исти проценат тачности може настати уз различит однос погрешних и недодељених исхода.

Издвојени тест скуп од 300 реченица погоднији је за формулисање главних закључака о успешности поступка, јер није коришћен за подешавање модела, промпта или инвентара након резултата прве итерације. Резултати за једночлане јединице на овом скупу приказани су у Табели 5.13.

Табела 5.13 Тачност модела на једночланим јединицама у издвојеном тест скупу од 300 реченица

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
base-lesk	47,2% (44,8-49,5)	50,0% (47,6-52,4)	+2,8
simple	48,2% (45,9-50,6)	50,6% (48,2-52,9)	+2,4

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
tesla	48,3% (45,9-50,6)	49,9% (47,5-52,2)	+1,6
multilingual-e5-large	52,6% (50,2-54,9)	56,0% (53,7-58,3)	+3,4
Llama4	68,7% (66,5-70,8)	75,3% (73,2-77,2)	+6,6
Llama4-ab1	65,1% (62,8-67,3)	69,7% (67,5-71,8)	+4,6
Llama4-ab2	68,9% (66,7-71,1)	75,3% (73,2-77,2)	+6,4
Llama4-ab3	64,6% (62,4-66,9)	70,8% (68,6-72,9)	+6,2
Mistral-Small-3.2	70,3% (68,1-72,4)	78,1% (76,1-80,0)	+7,8
GPT-3.5	62,4% (60,1-64,6)	-	-
GPT-4.1	79,2% (77,2-81,0)	92,4% (91,0-93,5)	+13,2
GPT-5.1	-	82,5% (80,7-84,2)	-
GeminiPro-2.5	73,1% (70,9-75,1)	81,5% (79,6-83,2)	+8,4

Табела 5.13 показује најјаснију разлику између породица модела. Полазни и енкодерски модели остају у зони од приближно 50-56% тачности у другој итерацији, док генеративни модели постижу знатно више вредности. Најбољи резултат остварује GPT-4.1 са 92,4% тачности, што је знатно изнад модела Llama4, Mistral-Small-3.2, GPT-5.1 и GeminiPro-2.5. Ово је најважнији табеларни резултат за једночлане јединице, јер је добијен на издвојеном тест скупу.

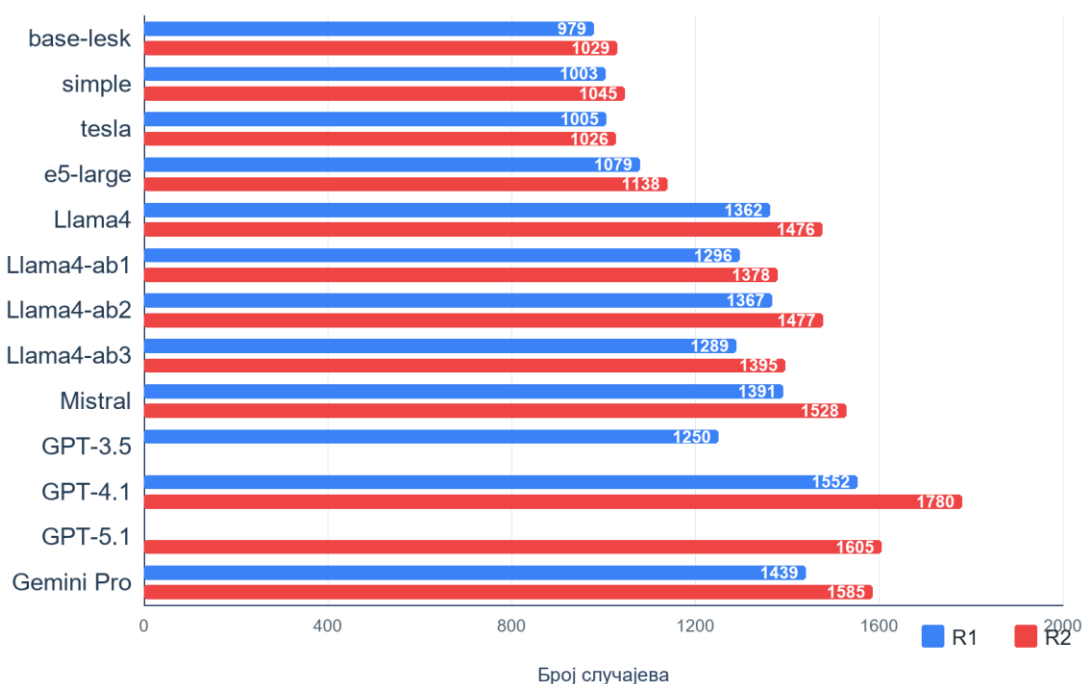
Табела 5.14 Тачност модела на вишечланим јединицама у издвојеном тест скупу од 300 реченица

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
base-lesk	86,1% (80,4-90,3)	86,6% (81,0-90,8)	+0,5
simple	89,3% (84,1-93,0)	89,3% (84,1-93,0)	0,0
tesla	89,8% (84,7-93,4)	89,8% (84,7-93,4)	0,0
multilingual-e5-large	89,3% (84,1-93,0)	86,1% (80,4-90,3)	-3,2
Llama4	91,4% (86,6-94,7)	91,4% (86,6-94,7)	0,0
Llama4-ab1	89,8% (84,7-93,4)	90,4% (85,3-93,8)	+0,6

Модел	R1 тачност (95% ИП)	R2 тачност (95% ИП)	Разлика R2-R1
Llama4-ab2	92,0% (87,2-95,1)	92,0% (87,2-95,1)	0,0
Llama4-ab3	89,8% (84,7-93,4)	89,8% (84,7-93,4)	0,0
Mistral-Small-3.2	92,0% (87,2-95,1)	92,5% (87,8-95,5)	+0,5
GPT-3.5	89,8% (84,7-93,4)	-	-
GPT-4.1	95,7% (91,8-97,8)	95,2% (91,1-97,4)	-0,5
GPT-5.1	-	93,0% (88,5-95,9)	-
GeminiPro-2.5	92,0% (87,2-95,1)	92,0% (87,2-95,1)	0,0

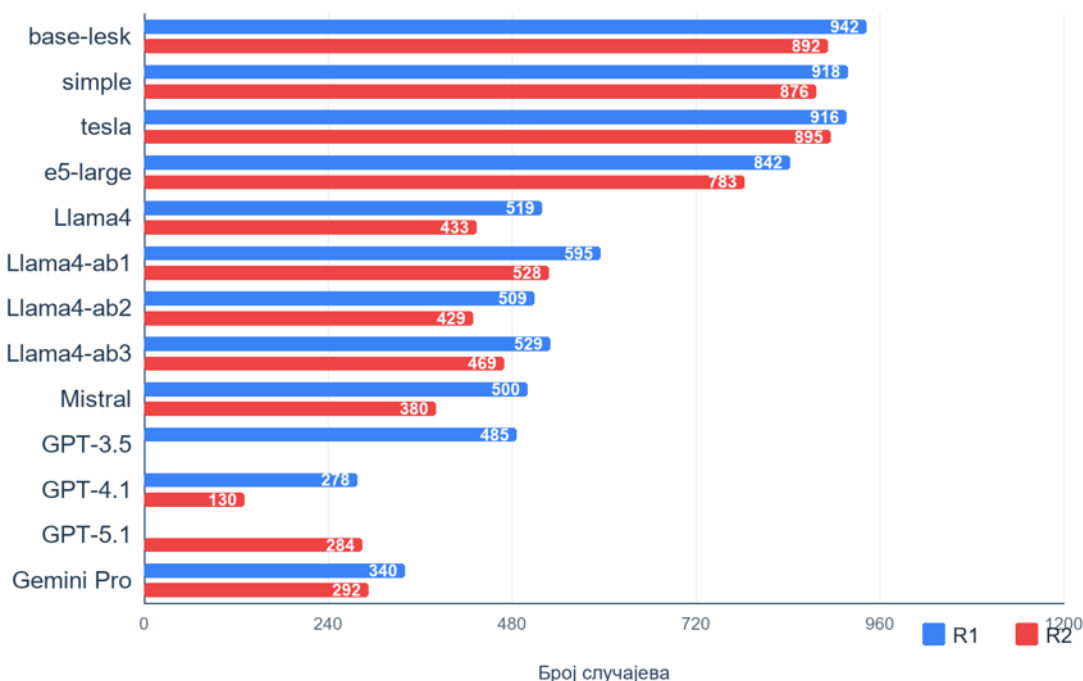
Табела 5.14 показује да су вишечлане јединице у целини лакше за обраду од једночланих. Разлика између прве и друге итерације овде је мања, а код неких модела готово је нема. Због тога главне закључке о напретку између итерација треба заснивати пре свега на резултатима за једночлане јединице из Табеле 5.13, док резултати из Табеле 5.14 потврђује да вишечлани изрази имају другачији профил тежине.

Као и код развојно-аналитичког скупа, Табеле 5.13 и 5.14 треба допунити графиконима расподеле исхода. Сlike 5.10, 5.11 и 5.11 приказују број тачних, погрешних и недодељених исхода на издвојеном тест скупу. На овим сликама се не уочава само који модел има већу тачност, већ већ оне указују и да ли модел више греша због избора погрешног значења или уздржавања од доделе. Осим тога, вредност овог скупа је што омогућава независнију проверу.



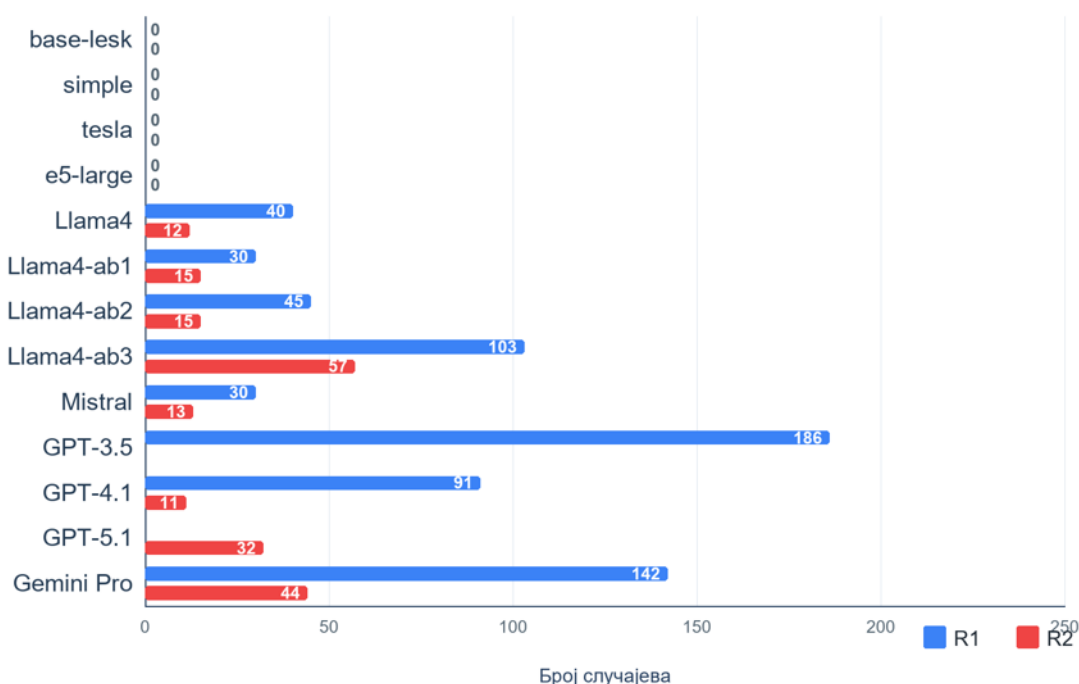
Слика 5.9 Број тачно разрешених случајева на издвојеном тест скупу од 300 реченица.

На издвојеном тест скупу, који није коришћен за накнадно подешавање поступка после прве итерације, раст броја тачних исхода остаје видљив. GPT-4.1 у R2 достиже 1780 тачних случајева од 1.921, у односу на 1552 у R1; следе GPT-5.1 са 1605 и Gemini Pro са 1585. Разлике између јачих генеративних модела постоје, али је још израженија разлика између њих и полазних или енкодерских поступака.



Слика 5.10. Број погрешно разрешених случајева на издвојеном тест скупу од 300 реченица.

Погрешно разрешени случајеви на тест скупу показују да друга итерација не уклања све грешке, али значајно смањује њихов број код најуспешнијих модела. Код GPT-4.1 број погрешних избора пада са 278 на 130, док код Mistral-Small-3.2 пада са 500 на 380. Насупрот томе, simple и tesla у R2 и даље имају значајан број погрешних исхода (876 односно 895), што показује да проширивање инвентара није довољно ако сам поступак слабије разликује блиска кандидатска значења.



Слика 5.11 Број недодељених случајева на издвојеном тест скупу од 300 реченица.

На тест скупу се такође види пад броја недодељених исхода у R2, али је тумачење опрезније јер овај скуп није био основ за развојну редакцију инвентара. GPT-4.1 смањује број недодељених случајева са 91 на 11, Llama4 са 40 на 12, а Gemini Pro са 142 на 44. Преостали недодељени исходи показују да и после редакције постоје случајеви у којима је потребна даља провера покривености, нормализације идентификатора⁵⁰ или разграничења понуђених значења.

Слике 5.9, 5.10 и 5.11 потврђују закључак из табела 5.13 и 5.14: GPT-4.1 у другој итерацији има највећи број тачних исхода и мали број недодељених случајева. Истовремено, поређење погрешних и недодељених исхода показује различите стратегије модела. Неки модели чешће бирају једно од понуђених значења и онда када избор није тачан, док други чешће остављају исход недодељеним. Та разлика је методолошки важна, јер недодељени исход може да сигнализира празнину у инвентару, док погрешан избор указује на проблем у разликовању постојећих кандидатских значења.

Додатна расподела по врстама речи показује да су глаголи најзахтевнија категорија, што је у складу са њиховом већом зависношћу од ширег синтаксичког и семантичког контекста.

Анализа у односу на број значења кандидата захтева опрезније тумачење. Код случајева са једним значењем кандидата не испитује се класично разрешење вишезначности, јер нема стварне конкуренције међу значењима. За Лесков поступак и енкодерске моделе то је углавном контролни случај који показује да ли је једино понуђено значење усклађено са златном ознаком. Код генеративних модела, међутим, тај случај није потпуно тривијалан, јер модел може да изабере NEW_SENSE, да произведе неусклађен идентификатор (изабере погрешано значење) или да одбије понуђено значење као неприкладно за контекст. Зато резултате за једно кандидатско значење треба поменути као проверу покривености инвентара и стабилности

⁵⁰ уједначавање начина на који су записане ознаке значења — на пример, уклањање сувишних размака како би се враћена ознака могла упоредити са ознаком у инвентару

излаза, али главни закључак о способности разрешења вишезначности треба заснивати на примерима са два или више кандидатских значења.

Резултати аблација показују да је пуна конфигурација промпта за *Llama4* у целини стабилнија од редуктованих варијанти. Уклањање захтева за образложењем углавном снижава тачност, што указује да образложење не служи само накнадном тумачењу резултата, већ помаже моделу да структурира одлуку о избору значења. Уклањање опције `NEW_SENSE` понекад повећава тачност, јер је модел приморан да изабере једно од понуђених значења, али се тиме губи важан сигнал за откривање недостајућих или недовољно разграничених значења у инвентару. Скраћени промпт остаје употребљив, али је мање стабилан. Због тога је за главни поступак оправдано задржати пуни промпт са образложењем и могућношћу избора `NEW_SENSE`, док резултате аблација треба тумачити као методолошку проверу, а не као посебну оптимизацију промпта. У целини, резултати показују да се највеће побољшање постиже у случајевима у којима је прва итерација била ограничена непотпуним или недовољно прецизним инвентаром значења. Због тога приказане вредности не треба тумачити само као поређење модела, већ и као емпиријски показатељ да итеративно унапређивање инвентара има непосредан утицај на квалитет аутоматске анотације значења.

5.7 Квалитативна анализа грешака и анализа вишезначних лема

Иако квантитативни резултати у табелама 5.13 и 5.14 показују јасно побољшање у другој итерацији, они не објашњавају у потпуности природу преосталих грешака. Због тога је, поред мерења тачности, спроведена и квалитативна анализа погрешно додељених и недодељених значења. Циљ ове анализе није само да се утврди који модел греша, већ да укаже на изворе грешака: развијене полисемије, блискост дефиниција, непотпуна покривености инвентара или недовољно јасно разграничења значења у самом ресурсу.

Квалитативна анализа мора се тумачити у односу на две различите улоге ручно проверених података. Развојно-аналитички скуп од 120 реченица, чија је структура приказана на Слици 5.3, служио је за анализу почетних резултата и редакцију инвентара. Издвојени тест скуп од 300 реченица, чија је структура приказана на Слици 5.4, служио је за процену текуће верзије поступка и за анализу преосталих грешака након редакције. Зато се грешке у првом скупу пре свега тумаче као основ за развој ресурса, док су грешке у другом скупу важније за процену ограничења поступка у текућој итерацији.

На развојно-аналитичком скупу посебно је видљив ефекат проширења инвентара значења. Друга верзија инвентара додала је 249 значења за 158 вишезначних лема: 140 лема проширено је за једно значење, 9 за два, 7 за три, док је лема „моћи“ проширена за пет нових значења. Та допуна у појединим случајевима непосредно побољшава резултат. На пример, лема „моћи“ је у почетној верзији имала само једно доступно значење, због чега модели у низу контекста нису могли да изаберу одговарајуће решење; после проширења инвентара, ти исти случајеви су се могли разрешити у другој итерацији.

Међутим, проширење инвентара не делује увек једносмерно. У неким случајевима оно отклања празнину у ресурсу, али истовремено повећава број значења кандидата и самим тим тежину избора. То се види код лема код којих се број значења кандидата повећава, али њихове дефиниције остају блиске или се делимично преклапају. Због тога резултате из табела 5.13 и 5.14 не треба тумачити само као поређење модела, већ и као показатељ међузависности модела и квалитета инвентара значења.

На издвојеном тест скупу анализа грешака има другачију улогу: она показује који проблеми остају и после редакције инвентара. Најпроблематичније леме у том скупу издвојене су у Табели 5.15. Стопа грешке представља удео погрешних одговора у укупном броју одговора модела за дату лему, при чему се одговори различитих модела на исту појаву речи броје засебно.

Табела 5.15 Најпроблематичније леме на издвојеном тест скупу

Лема	Опис тешкоће у разграничењу значења	Стопа грешке	Погрешно / укупно	Тумачење
земља	разграничење значења „држава”, „земља/тло” и блиских синсета	89,3%	150/168	Грешке упућују на блискост значења кандидата и потребу за јаснијим дефиницијама.
одређен	апстрактна и контекстуално зависна придевска значења	85,4%	82/96	Модел често мешају употребе у којима придев „одређен” означава неку, недовољно спецификовану јединицу или подскуп са употребама у којима означава прецизно утврђен, познат или прописан оквир.
добити	више значења у зависности од аргумената и ширег контекста	85,0%	102/120	Глаголска полисемија остаје један од најзахтевнијих делова задатка.
број	преклапање нумеричких, количинских и апстрактних употреба	81,9%	118/144	Висока учесталост не гарантује лако разрешење ако су значења функционално блиска.
моћи	модално и лексичко значење, уз више нових значења у v2	81,4%	469/576	Проширење инвентара помаже, али не уклања сложеност модалних и контекстуално зависних употреба.

Као што показује Табела 5.15, најзахтевније леме за разрешавање нису нужно оне које су ретке, већ оне код којих се у корпусу јавља више блиских значења осетљивих на контекст која је тешко разграничити. Насупрот томе, поједине учестале леме, као што су „година“ и „део“, имају ниже стопе грешке од најпроблематичнијих лема, што показује да сама учесталост није довољно објашњење. Тежина задатка пре свега зависи од броја кандидатских значења, грануларности инвентара и јасноће разграничења које је понуђено дефиницијама.

Посебно су важне грешке у којима тачно значење није било доступно у инвентару, а модел је ипак изабрао најближе понуђено значење. У таквим случајевима образложење модела често делује уверљиво на први поглед, али заправо показује да систем покушава на силу да смести појаву у најближи постојећи синсет. Овај тип грешке не указује само на слабост модела, већ пре свега представља сигнал да инвентар значења треба допунити, прецизније разграничити или другачије описати. Један пример такве грешке јавља се код придева „шпански“ у синтагми „шпански рат“. Уместо ширег релационог значења које се односи на Шпанију или на шпански народ, модел бира уже значење повезано са шпанским језиком. Сличан тип проблема јавља се

код глагола „осмислити“, где су два кандидатска значења у вези са планирањем толико блиска да је на основу саме дефиниције тешко поуздано раздвојити исправну интерпретацију. Ови примери показују да преостале грешке често настају у зони преклапања између језичког контекста и грануларности лексичко-семантичког ресурса.

Додатна анализа по врстама речи потврђује исти образац. На издвојеном тест скупу глаголи су најзахтевнија категорија: агрегирана тачност преко модела износи 57,8%, док је код придева 66,9%, код именица 72,3%, а код прилога 74,7%. Тај резултат је у складу са чињеницом да глаголска значења у већој мери зависе од аргументске структуре, ширег синтаксичког окружења и дискурсног контекста⁵¹. Због тога грешке код глагола треба тумачити као посебно важан показатељ да даље унапређење инвентара мора да обухвати не само додавање нових значења, већ и прецизне примере употребе.

И анализа у односу на број значења кандидата показује да додељивање исправног значења опада са бројем понуђених кандидата. На издвојеном тест скупу агрегирана тачност опада са 90,8% код јединица са једним значењем, на 71,8% код јединица са два значења, 62,5% код јединица са три значења и 40,2% код јединица са шест или више кандидатских значења. Овај налаз је у складу са расподелама приказаним у Табели 5.7 и Табели 5.8 и потврђује да број значења кандидата утиче не само на тежину избора који модел треба да учини, већ и на потребу за прецизнијим лексикографским разграничењем.

За разлику од броја кандидатских значења, дужина реченице није се показала као главни извор грешке. Тачност остаје релативно стабилна кроз различите групе реченичне дужине, што указује да проблем није пре свега у количини доступног контекста, већ у томе колико су кандидатска значења међусобно блиска и колико добро инвентар покрива стварну употребу у корпусу. Ово је важно јер усмерава наредне кораке: потребно је пре свега радити на редакцији значења и дефиниција, као и додавању примера, док је додавање дужег контекста у промпт у другом плану.

Коначно, недодељена значења и излази који се не могу поуздано повезати са инвентаром имају двоструку улогу. Са једне стране, они умањују тачност модела у евалуацији. Са друге стране, они су користан сигнал за развој ресурса, јер указујући на места на којима излаз није стабилно повезан са постојећим идентификаторима открива и потенцијалне недостатке инвентара. У другој итерацији број таквих случајева се смањује, што је у складу са проширењем инвентара, али они не нестају у потпуности.

У целини посматрано, квалитативна анализа показује да преостале грешке не треба сводити на слабост појединачног модела. Оне настају из међудејства три фактора: понашања модела, броја и блискости значења кандидата и покривености српског WSD инвентара. Управо зато анализа грешака има методолошку вредност за дисертацију: она показује да поступак разрешења вишезначности речи није само евалуација модела, већ и механизам за откривање празнина, преклапања и недовољно прецизних дефиниција у инвентару значења.

5.8 Резултати дестилације модела за разрешење вишезначности речи

У претходним одељцима приказани су резултати поступка у коме се разрешење вишезначности речи спроводи непосредном применом већих генеративних модела и ревидираног инвентара значења. Такав поступак има важну аналитичку улогу, јер омогућава да се уоче недостајућа значења, провери квалитет дефиниција и итеративно унапреди инвентара. Међутим, његова

⁵¹ Информације из претходних и наредних реченица релевантне за тумачење значења речи.

примена над већим корпусима може бити рачунски и организационо захтевна. Због тога је у посебном експерименту испитано у којој мери се понашање добијено полуаутоматском анотацијом може пренети на мањи модел, погоднији за локалну и бржу примену.

Дестилација је у овом експерименту схваћена у практичном смислу: мањи модел се не обучава на неким излазима наставничког модела, већ на сребрном скупу аутоматски означених примера, као што је наведено у одељку 3.4. Сваки пример се формулише као пар који чине реченица са означеном циљном јединицом и дефиниција једног кандидатског значења. Позитивни парови одговарају значењу додељеном у сребрној анотацији, док негативни парови обухватају друга кандидатска значења за исту лему и врсту речи. У фази примене модел бира оно кандидатско значење које најбоље одговара реченичном контексту.

Евалуација дестилованих модела спроведена је над истим издвојеним тест скупом који је коришћен у другој итерацији евалуације у претходном одељку. Скуп обухвата 1.921 аотирани сегмент текста: 1.734 једночлане и 187 вишечланих јединица. При поређењу предвиђених и референтних значења очувани су сви делови идентификатора који разликују појединачна значења. Када је циљна јединица имала само једно кандидатско значење, оно је додељивано директно, без избора модела између кандидата. Покривеност тест скупа износила је 1.894 од 1.921 примера, односно 98,6%. Преосталих 27 непокривених примера рачунато је као нетачно, а строга тачност израчуната је над свих 1.921 примера.

У експерименту су испитана три модела за избор кандидатског значења: *simple*, *tesla* и *multilingual-e5-large*. Детаљни параметри обуке нису неопходни у главној табели резултата, јер не мењају основни закључак о успешности дестилације. Зато је у Табели 5.16 приказано само поређење полазног модела и модела после дестилације. Као основна метрика коришћена је строга тачност над свим примерима у издвојеном тест скупу. То значи да се непокривени примери не изостављају из имениоца, већ се рачунају као нетачни.

Табела 5.16 Поређење полазних и дестилованих модела на издвојеном тест скупу

Модел	Полазни модел: тачно	Полазни модел: тачност	Дестиловани модел: тачно	Дестиловани модел: тачност	Разлика
simple	1 045 / 1 921	54,4%	1 315 / 1 921	68,5%	+14,1
tesla	1 026 / 1 921	53,4%	986 / 1 921	51,3%	-2,1
multilingual-e5-large	1 138 / 1 921	59,2%	1 394 / 1 921	72,6%	+13,3

Резултати показују да обука на сребрном скупу јасно побољшава два од три испитана модела. Модел *simple* побољшава тачност са 54,4% на 68,5%, што представља добитак од 14,1 процентних поена, односно 270 додатно тачних примера. Модел *multilingual-e5-large* побољшава тачност са 59,2% на 72,6%, што представља добитак од 13,3 процентна поена, односно 256 додатно тачних примера. Насупрот томе, модел *tesla* пада са 53,4% на 51,3%, па се у овој поставци не може сматрати успешним случајем дестилације.

Пошто су за полазне и дестиловане моделе доступне предикције над истим примерима, разлике су проверене Макнимаровим тестом, уз Холмову корекцију. Резултати су приказани у Табели

5.17. Табела је сведена на податке који су важни за тумачење: колико је случајева дестилација исправила, колико је нових грешака унела и да ли је промена статистички поуздана.

Табела 5.17 Парна провера разлике између полазних и дестилованих модела

Модел	Исправљено дестилацијом	Погоршано дестилацијом	Нето промена	Холмова р-вредност	Закључак
simple	425	155	+270	1,97e-29	дестиловани модел је значајно бољи
tesla	232	272	-40	8,22e-02	нема значајне разлике
multilingual-e5-large	435	179	+256	4,51e-25	дестиловани модел је значајно бољи

Ови резултати показују да је ефекат дестилације зависан од основног модела. Обука на сребрном скупу значајно побољшава моделе *simple* и *multilingual-e5-large*, али не доводи до поузданог побољшања модела *tesla*. Због тога је у даљем тумачењу важније пратити два успешна дестилована модела него све техничке варијанте обуке.

Поређење парова дестилованих модела приказано је у Табели 5.18. Ова табела служи само да покаже релативни поредак три дестилована модела, па су изостављене помоћне колоне које нису потребне за главни ток аргументације.

Табела 5.18 Поређење дестилованих модела

Поређење	Разлика тачности	Холмова р-вредност	Бољи модел
multilingual-e5-large / tesla	+21,5	5,89e-63	multilingual-e5-large
simple / tesla	+17,4	1,22e-42	simple
multilingual-e5-large / simple	+4,2	7,15e-07	multilingual-e5-large

Најбољи дестиловани модел је *multilingual-e5-large*, док је *simple* други по успешности. Иако је разлика између ова два модела мања него њихова разлика у односу на модел *tesla*, она је и даље статистички значајна (Макнимаров тест уз Холмову корекцију, $p = 7,15 \times 10^{-7}$). То оправдава избор модела *multilingual-e5-large* као главног дестилованог модела у даљем тумачењу резултата. Додатна анализа показује да добитак од дестилације није равномерно распоређен по свим типовима јединица. Да би табеле биле прегледније, једночлане и вишечлане јединице приказане су одвојено. Највећи део побољшања код модела *simple* и *multilingual-e5-large* долази од једночланих јединица, где је избор значења тежи и где је број значења кандидата обично већи.

Табела 5.19 Промена резултата код једночланих јединица

Модел	Полазни модел: тачно	Дестиловани модел: тачно	Нето промена
simple	877 / 1 734	1 144 / 1 734	+267
tesla	865 / 1 734	819 / 1 734	-46
multilingual-e5-large	971 / 1 734	1 221 / 1 734	+250

Табела 5.20 Промена резултата код вишечланих јединица

Модел	Полазни модел: тачно	Дестиловани модел: тачно	Нето промена
simple	168 / 187	171 / 187	+3
tesla	161 / 187	167 / 187	+6
multilingual-e5-large	167 / 187	173 / 187	+6

Табела 5.19 и Табела 5.20 показују да се главни добитак не јавља код вишечланих, већ код једночланих јединица. То је очекивано, јер вишечлане јединице већином имају специфичније значење и мањи број кандидатских значења, па је простор за побољшање мањи. Код једночланих јединица, напротив, дестилација значајно побољшава избор између конкурентних значења.

Посебно је важно истаћи да се добици код модела *simple* и *multilingual-e5-large* концентришу на случајеве где заиста постоји вишезначност. Примери са само једним значењем кандидата не представљају стварно разрешење двосмислености: у исправљеној евалуацији тај једини кандидат додељује се директно. Зато су они изостављени из наредних табела (Табела 5.21, Табела 5.22 и Табела 5.23), а приказ је ограничен на јединице са најмање два кандидатска значења. Мале разлике које се у радним резултатима могу јавити код случајева са једним кандидатом потичу од покривености и нормализације ознака, а не од способности модела да изабере одговарајуће значење.

Табела 5.21 Промена тачности према броју кандидатских значења: *simple*

Број кандидатских значења	Примери	Полазна тачност	Тачност после дестилације	Разлика
2	328	57,3%	64,9%	+7,6
3	260	41,5%	66,5%	+25,0
4–5	376	26,9%	51,1%	+24,2

Број кандидатских значења	Примери	Полазна тачност	Тачност после дестилације	Разлика
6–10	257	18,7%	39,3%	+20,6
11+	95	2,1%	40,0%	+37,9

Табела 5.22 Промена тачности према броју кандидатских значења: *multilingual-e5-large*

Број кандидатских значења	Примери	Полазна тачност	Тачност после дестилације	Разлика
2	328	61,6%	71,3%	+9,8
3	260	42,7%	71,2%	+28,5
4–5	376	41,0%	57,7%	+16,8
6–10	257	24,9%	45,1%	+20,2
11+	95	9,5%	45,3%	+35,8

Табела 5.23 Промена тачности према броју кандидатских значења: *tesla*

Број кандидатских значења	Примери	Полазна тачност	Тачност после дестилације	Разлика
2	328	54,0%	55,2%	+1,2
3	260	40,0%	32,7%	-7,3
4–5	376	26,6%	17,8%	-8,8
6–10	257	17,1%	19,8%	+2,7
11+	95	3,2%	6,3%	+3,2

Ова анализа потврђује да корист од дестилације не потиче само од једноставних случајева са једним кандидатом значења. Код модела *simple* и *multilingual-e5-large* побољшање се јавља у свим групама са два или више кандидата значења, а посебно је изражено у случају високо развијене полисемије. Модел *tesla* је нестабилан: има мала побољшања у неким групама, али се погоршава у групама са три, као и са четири и пет кандидатских значења.

На крају, дестиловане моделе треба поставити у однос са генеративним моделима из претходне евалуације, али без понављања целе табеле из претходног одељка. За овај одељак довољно је приказати најбољи дестиловани модел, основни генеративни модел *Llama4* и најуспешнији модел *GPT-4.1* (Табела 5.24). Аблационе варијанте промптова и додатни генеративни модели нису потребни у главној табели, јер су већ размотрени у претходном одељку.

Табела 5.24 Положај најбољег дестилованог модела у односу на генеративне моделе

Модел	Улога у поређењу	Тачно / укупно	Тачност
multilingual-e5-large, дестиловани модел	најбољи дестиловани модел	1 394 / 1 921	72,6%
Llama4	основни генеративни модел	1 477 / 1 921	76,9%
GPT-4.1	најбољи генеративни модел	1 784 / 1 921	92,9%

Најбољи дестиловани модел, *multilingual-e5-large*, достиже 72,6% тачности. Тај резултат заостаје за моделом *Llama4* за 4,3 процентна поена, а за моделом *GPT-4.1* за 20,3 процентна поена. Због тога дестиловане моделе не треба представити као замену за најуспешнији генеративни поступак, већ као оперативну допуну: они омогућавају бржу и јефтинију примену слоја за разрешење вишезначности речи, али уз јасан губитак тачности у односу на најбољи генеративни модел.

У целини, резултати показују да је обука на сребрном скупу корисна, али да њен ефекат зависи од основног модела. Најбољи резултат остварује *multilingual-e5-large*, док *simple* такође добија велико и статистички значајно побољшање. Модел *tesla* у овој поставци не показује поуздано побољшање. Главни добици јављају се код примера с једночланим лексичким јединицама које су вишезначне, што указује да дестилација побољшава способност модела да разликује конкурентна значења у контексту, а не само да решава тривијалне случајеве са једним кандидатом.

5.9 Резултати формирања српских синсета помоћу великих језичких модела

У овом експерименту испитана су три радна тока за формирање кандидата за српске синсете на основу енглеских синсета. Сваки радни ток производио је српску глосу и скуп српских литерала, али се разликовао по степену експлицитне семантичке обраде. Полазни радни ток представљао је директно превођење глосе и литерала. Вишефазни радни ток раздвајао је анализу синсета, превођење, проширивање и филтрирање литерала, као и завршно преформулисање глосе. Концептуални радни ток уводио је посредну проширену дефиницију као основу за избор литерала и коначно формулисање српске глосе.

Пуна серија генерисања обухватила је 1.734 редова са кандидатима, односно 578 редова у сваком од три радна тока. Као што показује Табела 5.25, радни извоз бележи 578 успешно обрађених радних јединица. Овај обим омогућава поређење три поступка на истом скупу улазних јединица, док се ручна евалуација у наставку посматра као мањи проверени подкуп пуне серије.

Табела 5.25 Обим генерисане серије и ручно провереног подскупа

Показатељ	Вредност
Укупно редова кандидата	1 734
Број јединствених енглеских синсета	578
Број радних токова	3
Ручно проверени редови са статусом „завршен“	90
Ручно проверени синсети	30

У пуној серији доминирају именички синсети, који чине 383 од 578 синсета. Остатак чине придеви, глаголи, придевски сателити (енгл. *adjective satellites*) и прилози. Расподела у Табели 5.26 важна је за тумачење резултата, јер задаци превођења глоса и избора литерала нису једнако захтевни за све врсте речи. Код глагола је посебно важно питање вида и односа свршених и несвршених глагола, код придева разликовање одређеног и неодређеног вида, док је код именица важна термилошка прецизност и доменска употреба.

Табела 5.26 Расподела пуне серије по врсти речи

Врста речи	Број радних јединица	Удео
именице (n)	383	66,3%
придеви (a)	59	10,2%
глаголи (v)	55	9,5%
придевски сателити (s)	54	9,3%
прилози (r)	27	4,7%
укупно	578	100,0%

Ручна експертска провера спроведена је на мањем подскупу од 30 синсета, при чему је за сваки синсет прегледан по један кандидат из сваког радног тока. Тако је добијено 90 проверених редова. Као што је приказано у Табели 5.27, у провереном подскупу налази се 12 именичких, 9 придевских, 5 глаголских и 4 прилошка синсета. Резултати зато имају карактер описне евалуације и не треба их тумачити као статистички довољну основу за коначни избор поступка на нивоу целог ресурса. Евалуатор је квалитет глосе по разним критеријумима, као и квалитет литерала оцењивао оценама од 1 до 5.

Табела 5.27 Расподела ручно провереног подскупа

Врста речи	Број синсета	Број редова кандидата
именице (n)	12	36
придеви (a)	9	27
глаголи (v)	5	15
прилози (r)	4	12
укупно	30	90

Просечне оцене глоса у Табели 5.28 показују да сва три радна тока производе српске дефиниције које су у великом броју случајева формулисане у духу српског језика и формално усклађене са очекиваном структуром глосе ворднета. Истовремено, полазни радни ток има најнижи просек од пет критеријума за глосу, што је очекиван резултат с обзиром на то да се ослања само на превођење без укључивања додатне семантичке стабилизације. Концептуални радни ток има највиши просек, 4,71, вишефазни радни ток 4,56, док је просек полазног радног тока 4,36. Овај однос подржава методолошку тврдњу да за формирање употребљивог српског синсета није довољан минимални преводни контекст, већ је потребан поступак који експлицитније контролише значење пре коначне формулације.

Табела 5.28 Просечне оцене по радном току

Критеријум	Полазни радни ток	Вишефазни радни ток	Концептуални радни ток
Природност глосе	4,33	4,60	4,72
Семантичка адекватност глосе	4,43	3,97	4,45
Разликовање значења	4,31	4,41	4,57
Структура глосе према врсти речи	4,30	5,00	5,00
Циркуларност глосе	4,40	4,87	4,79
Просек пет критеријума за глосу	4,36	4,56	4,71
Адекватност литерала	4,10	3,29	4,10

Предности и ограничења разматраних радних токова испољавају се у различитим аспектима. Концептуални радни ток има највише оцене за природност глосе, разликовање значења и укупни просек глосе. Вишефазни радни ток има врло добре оцене за структуру глосе према врсти речи и избегавање циркуларности, али је његова просечна оцена семантичке адекватности глосе нижа него код друга два приступа. То показује да формално добро организован поступак не гарантује увек најпрецизније очување значења. Полазни радни ток је, с друге стране, једноставнији и у неким случајевима даје сасвим употребљиве резултате, али даје најслабије резултате у формулацији глоса, нарочито по критеријумима који зависе од лексикографске контроле, као што су структура дефиниције и избегавање циркуларности.

Оцене литерала показују другачију слику од оцена глоса. Полазни и концептуални радни ток имају исту просечну оцену адекватности литерала, 4,10, док је вишефазни радни ток знатно слабији, са просеком 3,29. Овај резултат упућује на то да се проширивање и филтрирање кандидата за литерале не решава само додавањем већег броја могућих превода. Напротив, управо вишефазни ток показује да шире и мање рестриктивно проширивање скупа кандидата може увести сувишне, семантички померене, непотврђене⁵² или неприхватљиво генерисане

⁵² Примери у Табели 5.30 показују да проширивање скупа кандидата може довести до укључивања неприхватљивих литерала. Тако је за синсет *спуу.у.01* вишефазни радни ток, поред одговарајућег литерала „завидети“, предложио и неприхватљив облик „завидимити“. Овај пример показује да повећање броја

литерале. За запис из ворднета то је озбиљан проблем, јер синсет не представља листу приближних преводних еквивалената, већ скуп лексема које деле исто значење.

Показатељи из ручне провере у Табели 5.29 потврђују да добијени кандидати не могу бити непосредно унети у Српски ворднет без ручне редакције. У већини редова означена је потреба за већом или мањом исправком глосе. Број глоса без потребне исправке највећи је код концептуалног радног тока, али је и ту реч о малом броју проверених редова. Посебно је значајно да су недостајући литерали забележени у више од половине проверених редова у сва три радна тока. Сувишни литерали најизраженији су у вишефазном радном току, где су забележени у 26 од 29 попуњених случајева. То је најјачи квантитативни показатељ да вишефазни ток, иако сложенији од полазног, није нужно бољи од концептуалног приступа. Наиме, у вишефазном радном току сувишни литерали забележени су у 26 од 29 случајева за које је овај показатељ попуњен. Њихова висока учесталост у ручно провереном подскупу указује на потребу за строжом провером припадности предложених литерала значењу синсета.

Табела 5.29 Показатељи из ручне провере

Показатељ	Полазни радни ток	Вишефазни радни ток	Концептуални радни ток
Исправка глосе није потребна	2/30	1/30	4/30
Потребна мања исправка глосе	11/30	9/30	11/30
Потребна већа исправка глосе	17/30	20/30	12/30
Ознака исправке глосе није попуњена	0/30	0/30	3/30
Забележени сувишни литерали	13/30	26/29	17/30
Забележени недостајући литерали	18/30	17/30	18/30
Нема ни сувишних литерала ни литерала који недостају	8/30	1/30	6/30

Квалитативни примери у Табели 5.30 показују шта ови бројеви значе у конкретним случајевима. Посебно су важни примери у којима је текст на први поглед у духу српског језика али је ипак лексикографски неприкладан, јер припада ширем, ужем или блиском, али ипак различитом, значењу. Такви случајеви показују да се евалуација не може свести на проверу природност генерисаног српског текста.

предложених превода захтева проверу језичке исправности сваког кандидата, као и његове припадности значењу синсета. Синсет чине лексичке јединице које деле исто значење.

Табела 5.30 Илустративни примери из ручне провере

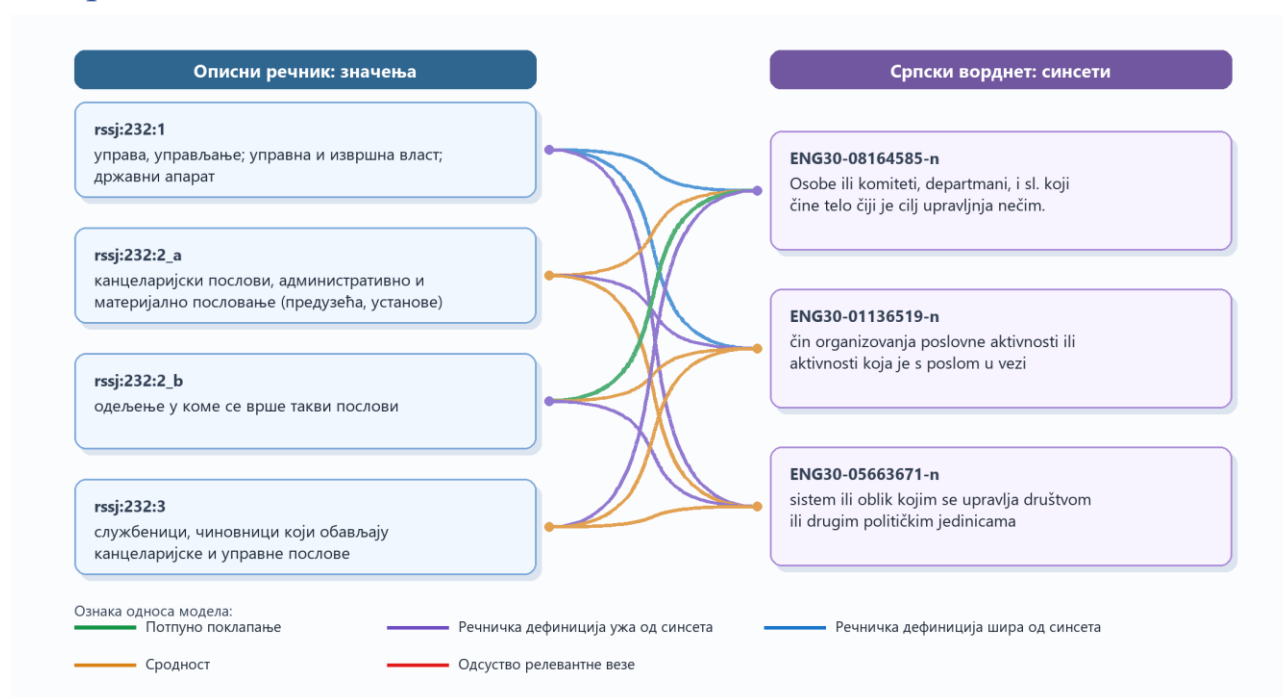
Синсет	Радни ток	Пример излаза	Шта пример показује
musical_performance.n.01	вишефазни	извођење; изведба музике; музички наступ; музички перформанс; музичирање; музички концерт	Прекомерно проширивање скупа литерала; анотатор бележи да је више предлога неадекватно, док је погодније <i>музичко извођење</i> или <i>извођење музике</i> .
envy.v.01	вишефазни	завидети; завидимити	Облик <i>завидимити</i> представља јасан пример непостојећег или неприхватљивог литерала који је настао при аутоматском генерисању.
hairless.a.01	концептуални	бездлак; бездлакаст	И сложенији ток може произвести сумњив литерал; у коментару је забележено да <i>бездлакаст</i> делује измишљено.
discouraging.a.01	полазни	одвратно; одстрашујуће	Полазни ток може промашити значење непосредним превођењем и понудити литерале који не припадају датом синсету - <i>одвратно</i> . У овом случају понуђен је и „измишљени“ литерал <i>одстраићујуће</i>
osmosis.n.01	вишефазни	osmosis	Задржавање енглеског облика уместо српског <i>осмоза</i> показује проблем неконтролисаног преноса изворног језика.
encouraging.a.01	вишефазни	мотивациони; инспиративан; подижући	Додати литерали су речи српског језика, али не могу се наћи у истом синсету.
drop.v.03	сви токови	пасти, опати, смањити се, спустити се	Пример показује да је код глагола важно питање вида и семантичког обима, које се не може поуздано решити дословним превођењем.

За дискусију су посебно корисна два типа примера из Табеле 5.30. Први су примери у којима је грешка очигледна, као што је *завидимити*, јер они показују да аутоматски поступак може да произведе непостојећи облик. Други примери су суптилнији, као што је *музички наступ* за *musical performance.n.01*: понуђени кандидат није потпуно погрешан, али је у односу на синсет за који је предложен ужи и више везан за сценски наступ, док су *музичко извођење* и

извођење музике прикладнији. Управо такви случајеви показују да евалуација мора бити лексичко-семантичка, а не само преводилачка.

У целини, резултати показују да велики језички модели могу да произведу релативно квалитетне кандидате за српске глосе, нарочито када је поступак организован тако да се значење синсета експлицитно стабилизује пре коначне формулације. Полазни ток је најслабији на нивоу глоса, што је у складу са очекивањем да непосредно превођење није довољно за лексикографски употребљив запис. Међу сложенијим поступцима концептуални ток је стабилнији од вишефазног, пре свега зато што вишефазни ток има изражену склоност ка додавању сувишних и неприкладних литерала. Истовремено, резултати за литерале и број потребних лексикографских исправки показују да аутоматски добијен излаз треба третирати као полазни материјал за рад на проширивању Српског ворднета, а не као завршен лексикографски запис.

5.10 Резултати поравнања лексичко-семантичких ресурса са описним речником



Слика 5.12 Пример поравнања значења из описног речника и синсета Српског ворднета за лему „администрација”.

У овом одељку приказани су резултати поравнања речничких значења из описног речника са кандидатским синсетима Српског ворднета. Слика 5.12 приказује један пример таквог поравнања. Поступак није постављен као потпуно аутоматско уношење нових веза у ресурс, већ као контролисано издвајање и класификација парова кандидата. За сваки пар речничког значења и синсета модел је додељивао једну од пет ознака: потпуно поклапање, речничка дефиниција ужа од синсета, речничка дефиниција шири од синсета, сродност значења или одсуство релевантне везе. Поред ознаке односа, модел је давао и процену сопствене поузданости и кратко образложење, што је омогућило да се резултати анализирају не само квантитативно, већ и као материјал за лексикографску редакцију.

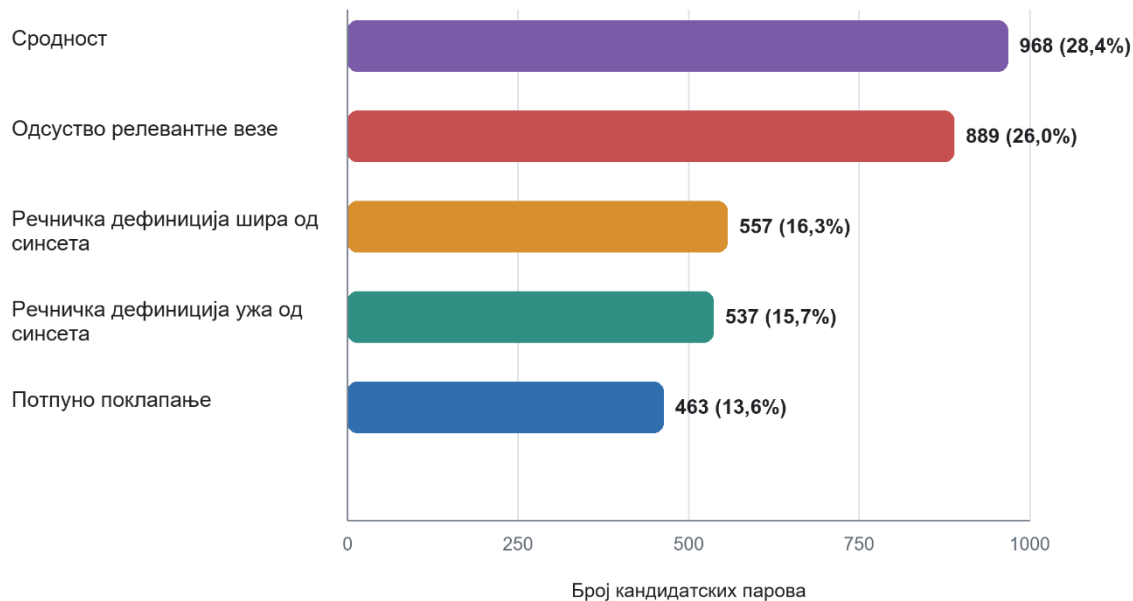
Поступак поравнања треба разликовати на три нивоа: ниво парова кандидата, ниво речничких значења и ниво лема. Улазни РССЈ скуп обухвата 1 829 речничких значења повезаних са 535 одредница, од чега 1.799 значења има придружен неки дефиницијски текст. На основу заједничке леме и компатибилне врсте речи формиран је скуп од 3 414 парова кандидата између речничких значења и синсета Српског ворднета. У парове кандидата ушла су 1 592 речничка значења, док за 207 значења са истом придруженом дефиницијом није пронађен одговарајући кандидатски синсет.

Прве четири ознаке (потпуно поклапање, речничко значење уже од синсета, речничко значење шире од синсета, сродан однос) у овој анализи третиране су као позитивне или потенцијално позитивне везе, док ознака одсуства релевантне везе означава кандидате које треба одбацити. На нивоу речничких значења, 1 422 значења имала су најмање један пар кандидата коме је модел доделио позитивну ознаку, док је 170 значења имало само кандидате за које је модел установио да не постоји релевантна веза. На нивоу лема, парове кандидата имале су 463 леме; од тога је 458 лема од модела добило бар једну позитивну ознаку модела, а 5 лема само кандидате без релевантне везе. Додатних 67 лема из улазног скупа није било обухваћено паровима кандидата. Основни показатељи обухвата приказани су у Табели 5.31.

Табела 5.31 Основни показатељи скупа за поравнање

Показатељ	Број	Удео / напомена
Полазна речничка значења	1 799	100,0%
Речничка значења са најмање једним потенцијалним синсетом	1 592	88,5% полазног скупа
Речничка значења без потенцијалног синсета	207	11,5% полазног скупа
Речничка значења са најмање једном позитивном ознаком модела	1 422	79,0% полазног скупа; 89,3% значења са кандидатима
Речничка значења са кандидатима, али без позитивне ознаке модела	170	9,4% полазног скупа; 10,7% значења са кандидатима
Парови кандидата	3 414	сви парови дефиниција–синсет
Потенцијални синсети	768	синсети који се јављају у излазу модела
Позитивне ознаке додељене моделом	2 525	74,0% парова кандидата

Расподела ознака које је дао модел, приказана на Слици 5.13, показује да поравнање није бинарни задатак у коме већина парова једноставно припада категорији потпуног поклапања или непоклапања. Најчешћа позитивна ознака је сродност, са 968 парова, док потпуно поклапање чини 463 пара, односно 13,6% свих парова кандидата. Ова расподела је важна јер указује на то да описни речник и Српски ворднет често захватају исти или близак простор значења, али га не сегментирају на исти начин.



Слика 5.13 Расподела ознака које је доделио модел у пуном скупу

Додатни показатељи, наведени у Табели 5.32, потврђују да су разлике у грануларности један од централних проблема овог поравнања. Укупно 563 речничка значења имају више позитивно означених синсета, што значи да једна речничка дефиниција може обухватити више јединица у ворднету или да модел препознаје више блиских кандидата. Са друге стране, 665 кандидата за синсете има више позитивно означених речничких значења, што указује на обрнуту ситуацију: више речничких дефиниција може бити доведено у везу са истим синсетом. Ови резултати не морају нужно представљати грешку модела при процени семантичке везе између речничког значења и синсета, већ често одражавају различите принципе организације значења у речнику и у ворднету.

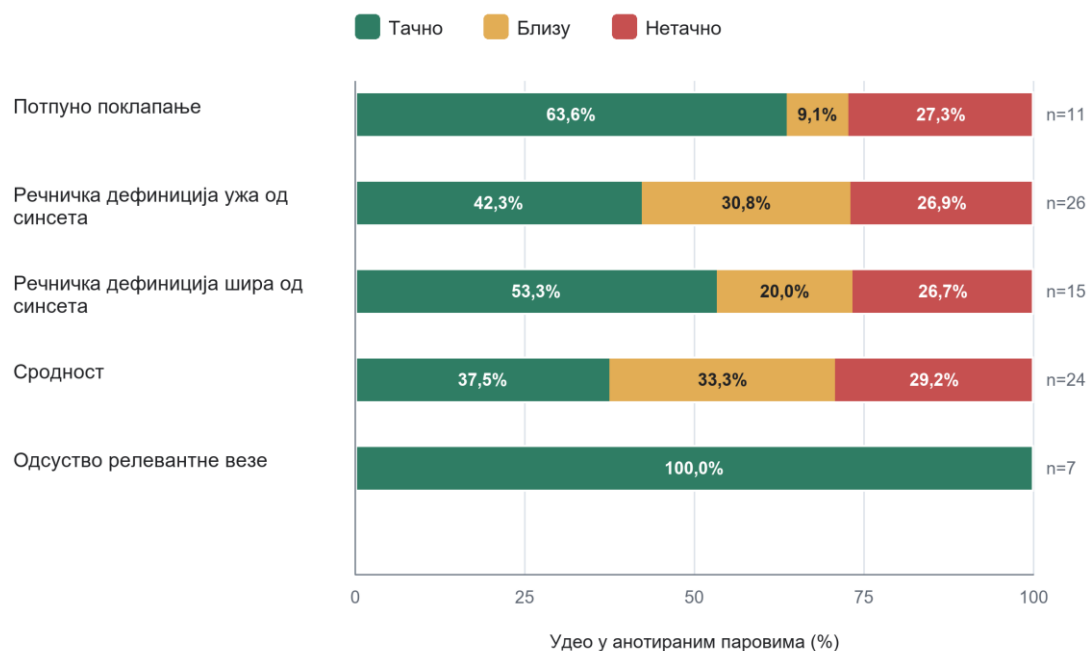
Табела 5.32 Показатељи разлика у грануларности речничких значења и синсета

Показатељ	Број	Тумачење
Речничка значења са више позитивно означених синсета	563	једна речничка дефиниција може покривати више синсета или више блиских кандидата
Кандидатски синсети са више позитивно означених речничких значења	665	више речничких значења може одговарати истом или блиском синсету
Кандидатски синсети којима модел није доделио позитвну ознаку	27	синсети који су се јавили као кандидати, али нису прихваћени ни у једном пару

Ручна провера је спроведена над првим пакетом од 100 парова кандидата. Тај пакет обухвата 22 леме, 59 речничких значења и 35 синсета. Анотирано је 83 пара, док је 17 парова остало необрађено. Потпуно је проверено 19 лема: „абдицирати”, „аберација”, „абразија”, „адаптација”, „адаптирати”, „адаптирати се”, „адхезија”, „администрација”, „администратор”,

„адреса”, „адресар”, „адресирати”, „адвокат”, „аеросол”, „афирмативан”, „агенција”, „агент”, „агитација” и „агломерација”. Три леме из пакета нису биле обрађене у ручној провери: „агенс”, „агресиван” и „аглутинација”.

У ручно провереном подскупу⁵³, 42 од 83 анотирана пара оцењена су као тачна, 20 као блиска, а 21 као нетачан. Ако се као условно прихватљиви рачунају и тачни и блиски случајеви, прихватљивост износи 74,7%. Ову вредност не треба тумачити као коначну тачност целог система, јер је проверен само један мањи пакет, већ као показатељ да излаз модела садржи значајан број употребљивих кандидата, али и довољно грешака, због чега је ручна провера неопходна. Резултати ручне оцене по ознаци модела приказани су на Слици 5.14.

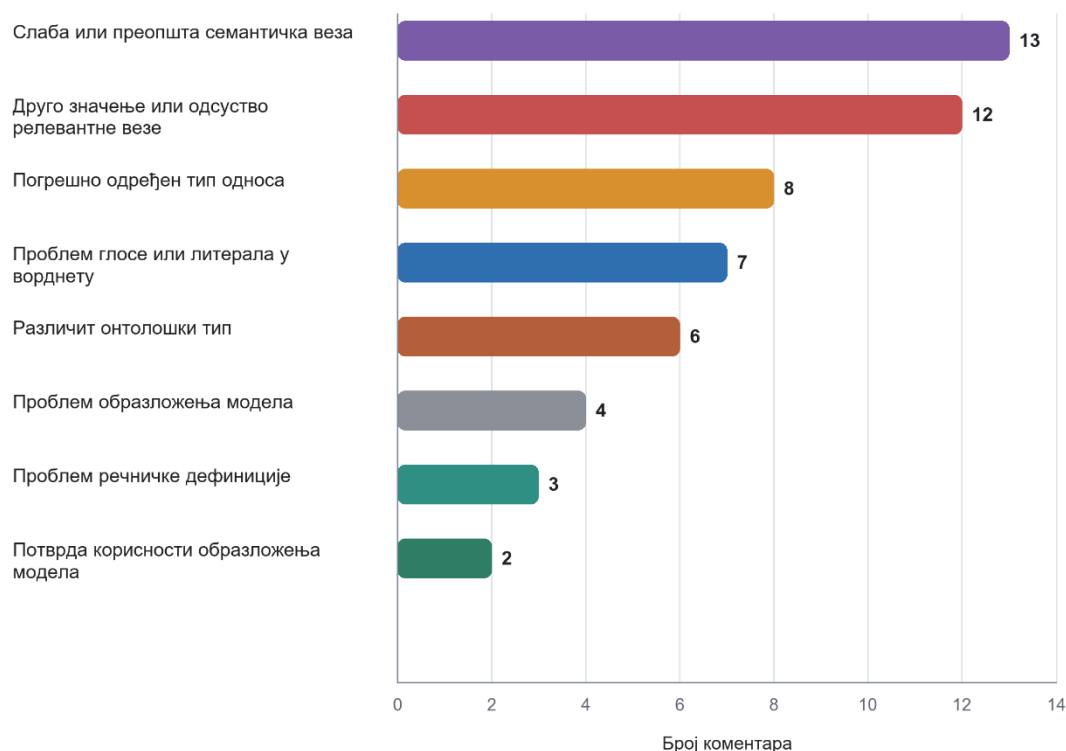


Слика 5.14 Ручна оцена ознака које је доделио модел у провереном пакету

Сви случајеви одсуства релевантне везе су се показали као тачни, али је та категорија у провереном пакету заступљена малим бројем анотираних примера. Код позитивних ознака уочава се да све ознаке имају сличан удео тачних и блиских оцена, али се разликују по природи грешака. Ознаку сродности је анотатор најчешће означавао као нетачну, што је очекивано, јер је реч о најслабије ограниченој позитивној категорији. Управо код сродности граница између корисног кандидата и сувише опште асоцијације постаје најосетљивија.

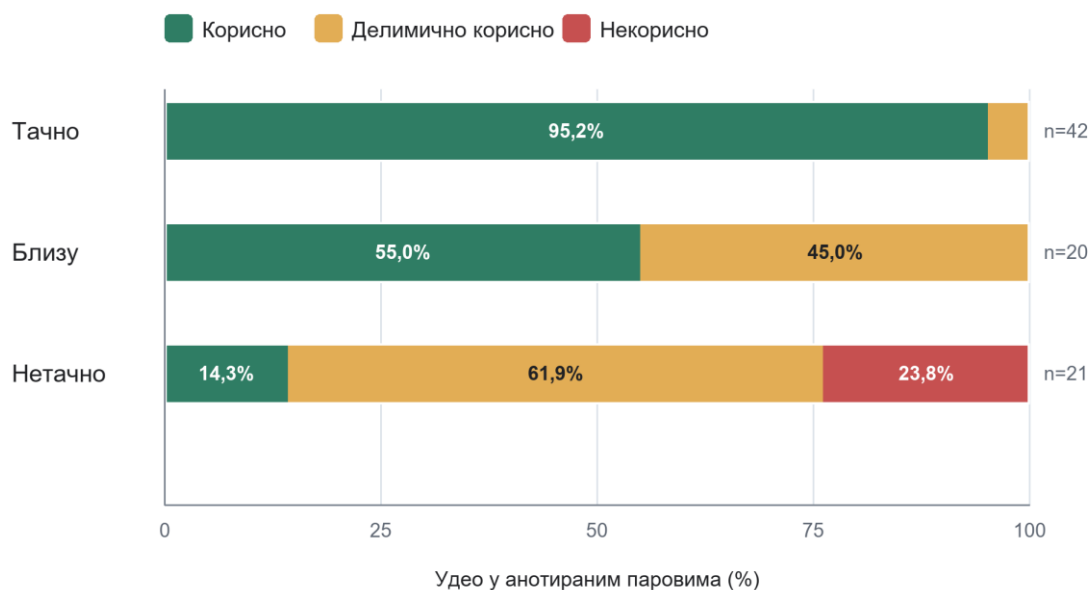
Коментари анотатора омогућили су додатну класификацију типова проблема. Најчешће се јављају слаба или преопшта семантичка веза (13 коментара), друго значење или одсуство релевантне везе (12 коментара), и погрешно одређен тип односа (8 коментара). Као што је приказано на Слици 5.15, највећи део спорних случајева није последица техничке грешке у формату излаза, већ семантичке неодређености, различите грануларности или погрешно изабраног смера односа.

⁵³ Анотатор је исправност ознаке односа коју је предложио модел оцењивао категоријама „тачно”, „близу” и „нетачно”. Посебно је бележио колико је сигуран у сопствену оцену, на скали од 1 до 5, при чему већа вредност означава већу сигурност. Модел је уз предложену ознаку односа исказивао степен сигурности у распону од 0 до 1. У даљем тексту ове две величине називају се сигурност анотатора у сопствену оцену и исказана сигурност модела.



Слика 5.15 Класе коментара анотатора

Корисност образложења које даје модел представља посебно важан резултат. У 54 од 83 аотирана случаја, односно у 65,1% проверених парова, образложење је оцењено као корисно. У још 24 случаја, односно 28,9%, оно је оцењено као делимично корисно, док је потпуно некорисних образложења било мало: 5 случајева, односно 6,0%. Одатле следи да је главни налаз да је већина образложења корисна, значајан део је делимично користан, а само мањи број образложења није помогао у редакцији. Однос између ручне оцено и корисности образложења приказан је на Слици 5.16. Ови подаци не значе да су образложења доказ исправности ознаке коју је дао модел, већ да су корисна за процес људске редакције: она показују на основу чега је модел повезао две јединице и тако могу да помогну анотатору да брже прихвати, коригује или одбаци предложени однос.



Слика 5.16 Корисност образложења које је понудио модел према оцени аотатора

У подскупу од 83 оцењена пара утврђена је позитивна корелација између исказане сигурности модела и сигурности аотатора у сопствену оцелу: Пирсонов (Pearson) коефицијент износи 0,2939, а Сперманов (Spearman) 0,3182. Ови коефицијенти описују повезаност двеју процена сигурности. За тумачење резултата потребно је размотрити и како је аотатор оцелио исправност предложених односа.

У 26 парова аотатор је сопствену сигурност оцелио вредношћу 3; ознаке модела у тим паровима оцењене су као тачне у три случаја, блиске у 12 и нетачне у 11. Међу тих 26 парова, модел је исказао сигурност нижу од 0,90 у девет случајева, а најмање 0,90 у преосталих 17. У целом подскупу од 83 пара сигурност модела била је нижа од 0,90 у 22 случаја. Истовремено, девет ознака за које је модел исказао сигурност од најмање 0,92 аотатор је оцелио као нетачне. Ови случајеви указују на то да висока исказана сигурност модела сама по себи није довољна за прихватање предложеног односа без ручне провере.

За дискусију су посебно важни случајеви у којима је модел имао високо поверење у своју одлуку, али је ручна оцена била ниска или негативна. Они показују да модел често препознаје површински или општи семантички траг, али га прецењује као довољан основ за поравнање. На пример, код леме „адаптација“ модел повезује значење „прилагођавање стране речи“ са синсетом који описује дело прерађено у други облик, јер обе јединице деле општи појам адаптације. Коментар аотатора, међутим, указује да је таква веза сувише општа. Слично томе, код леме „адхезија“ речничка дефиниција описује силу, док глоса синсета описује својство или приањање површина, па проблем није само у блискости термина, већ у различитом онтолошком типу. Код леме „администратор“ веза постоји, али је модел погрешно одређује као однос ширег и ужег значења, док је ручна процена да је реч пре о сродним значењима.

У целини, резултати показују да генеративни модел може да издвоји велики број релевантних потенцијалних веза између речничких дефиниција и синсета, али њих треба тумачити као кандидате за полуаутоматско поравнање. Поступак је нарочито користан за откривање потенцијалних поклапања, сродних значења и разлика у грануларности, као и за идентификацију проблематичних места у оба ресурса. Истовремено, ручна провера показује да високо поверење модела у свој избор и уверљиво образложење нису довољни за аутоматско

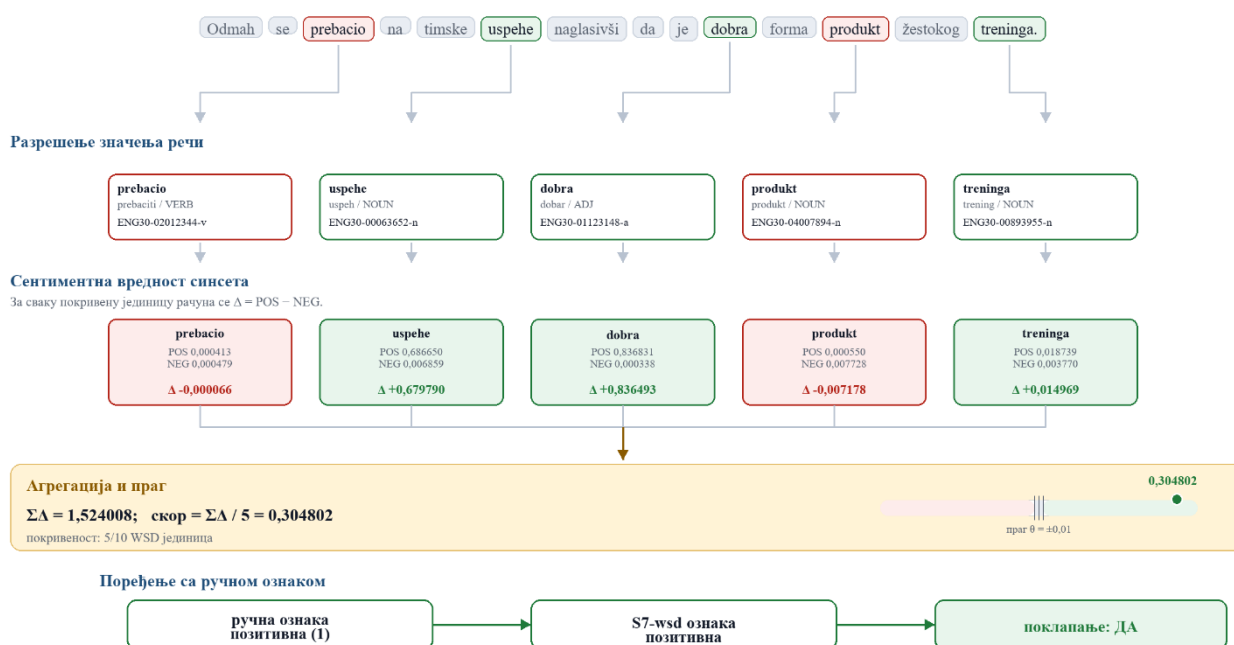
прихватање поравнања. Коначна одлука мора остати везана за експертску процену односа између речничке дефиниције, глосе синсета и места синсета у широј лексичко-семантичкој структури.

5.11 Екстерна евалуација обогаћеног Српског ворднета у анализи сентимента

Екстерна евалуација обогаћеног Српског ворднета спроведена је на задатку анализе сентимента реченица. За разлику од интерних провера у одељцима 0, 5.2.4, ова евалуација не пореди лексиконе само са глосама синсета, већ испитује да ли вредности сентимента на нивоу синсета доводе до бољег предвиђања поларитета целе реченице. Тиме се проверава практична употребљивост добијеног слоја сентимента у низводном задатку.

Реченица

Обојене речи су WSD јединице које имају вредност у S7; сиве речи не улазе у овај просек.



Слика 5.17 Пример израчунавања сентимента реченице у S7-wsd конфигурацији

Евалуација је спроведена над 2.536 ручно аотираних реченица из полазног узорка од 3.000 новинских реченица, описаног у одељку 3.7. У скупу је било 908 негативних, 989 неутралних и 639 позитивних реченица. За сваку реченицу израчуната је просечна вредност сентимента на основу оних јединица за које је одговарајући ресурс обезбедио вредност сентимента на нивоу лексеме или на нивоу синсета. Коначна ознака реченице добијена је поређењем те вредности са прагом θ . Слика 5.17 приказује један пример, а поступак је детаљније описан у одељку 3.7. За сваки лексикон у табели (Табела 5.33) приказан је праг који је дао најбољу макро-F1 меру на евалуационом скупу.

Табела 5.33 Резултати екстерне евалуације у анализи сентимента реченица

Конфигурација	Поступак	Праг θ	Тачност	Макро-F1	Просечна покривеност
---------------	----------	---------------	---------	----------	----------------------

S7-avg	просек по лема и врсти речи	0,02	0,4543	0,4510	0,6272
S7-wsd	изабрани WSD синсет	0,01	0,4550	0,4441	0,3956
S6-avg	просек по лема и врсти речи	0,02	0,4519	0,4440	0,6272
S5-avg	просек по лема и врсти речи	0,03	0,4476	0,4417	0,6272
S2-avg	просек по лема и врсти речи	0,02	0,4432	0,4324	0,6272
S5-wsd	изабрани WSD синсет	0,01	0,4314	0,4277	0,3956
L0	лексичке вредности лема	0,00	0,4227	0,4208	0,2360
S4-wsd	изабрани WSD синсет	0,01	0,4310	0,4161	0,3956
S4-avg	просек по лема и врсти речи	0,03	0,4278	0,4159	0,6272
S6-wsd	изабрани WSD синсет	0,01	0,4405	0,4065	0,3956
S2-wsd	изабрани WSD синсет	0,00	0,4267	0,3981	0,3956
S3-avg	просек по лема и врсти речи	0,01	0,4152	0,3940	0,6272
S0	полазно SWN пресликавање и WSD синсет	0,00	0,4097	0,3924	0,3956
S3-wsd	изабрани WSD синсет	0,01	0,4073	0,3788	0,3956
S1-wsd	изабрани WSD синсет	0,14	0,3955	0,3148	0,3956
S1-avg	просек по лема и врсти речи	0,20	0,3691	0,2825	0,6272

Резултати показују да најбољи учинак по макро-F1 мери постиже конфигурација S7-avg, заснована на просеку вредности синсета за исту лему и врсту речи у лексикону S7. Њена тачност износи 0,4543, макро-F1 мера 0,4510, а просечна покривеност 0,6272. Врло близак резултат има S7-wsd, конфигурација која користи синсет изабран поступком разрешења вишезначности речи: она има нешто вишу тачност (0,4550), али нешто нижу макро-F1 меру (0,4441) и знатно нижу покривеност (0,3956). Овај однос је методолошки важан. Конфигурација avg не решава значење појединачне речи у контексту, већ користи просечан

сигнал за дату лему и врсту речи. Њена предност у покривености указује да тренутни инвентар значења и његова повезаност са слојем сентимента још увек ограничавају пуну употребу приступа заснованог на разрешењу вишезначности речи. Укупно је у WSD фази било 29.284 релевантних циљних јединица; кандидати су пронађени за 19.496 јединица, изабрани синсет је имао доступну вредност сентимента у 11.636 случајева, док је бар један кандидат за одговарајући синсет имао вредност сентимента у 14.663 случаја. То показује да део потенцијалног добитка остаје ограничен покривеношћу и усклађеношћу ресурса, а не нужно квалитетом самих вредности сентимента.

Полазни лексикон S0, заснован на директном пресликавању из енглеског SentiWordNet-a, има слабији резултат од најбољих обогачених конфигурација. Његова тачност износи 0,4097, а макро-F1 мера 0,3924. У поређењу са њим, S7-avg повећава макро-F1 за 0,0586, док S7-wsd повећава макро-F1 за 0,0516. Овај налаз је у складу са резултатима из одељка 5.2.3: лексикони настали обучавањем и прилагођавањем на српским подацима боље одражавају односе сентимента у српском језику него директно пресликавање из енглеског ресурса.

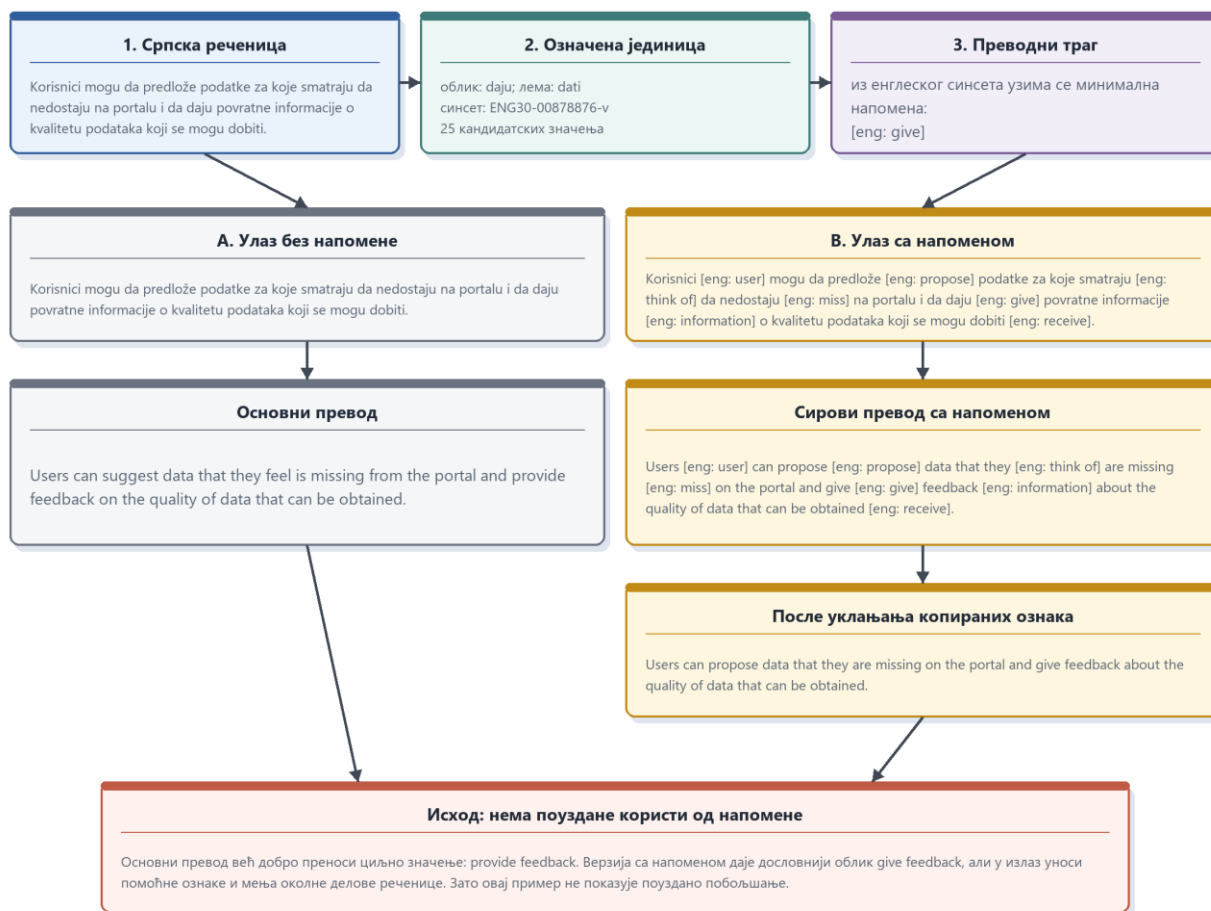
Табела 5.34 Тестови статистичке значајности за главна поређења

Поређење	Δ тачности	Макним аров тест, р	Δ макро- F1	Рандомиз ација р	Холмова корекција , р	Значајно на 0,05
S7-avg према L0	0,0315	0,0123	0,0302	0,0175	0,0525	не
S7-avg према S0	0,0446	0,0003	0,0586	0,0005	0,0025	да
S7-wsd према S0	0,0453	0,0000	0,0516	0,0005	0,0025	да
S7-avg према S7-wsd	-0,0008	0,9710	0,0070	0,5107	1,0000	не
S7-avg према S6-avg	0,0024	0,8452	0,0070	0,5122	1,0000	не

За тачност је примењен Макнимаров тест над упареним предвиђањима сентимента реченица, а за макро-F1 меру приближна рандомизација над истим упареним предвиђањима, а резултати су приказани у Табели 5.34. Пошто је испитано више унапред дефинисаних поређења, вредности за макро-F1 кориговане су Холм-Бонферонијевим поступком. После корекције, и S7-avg и S7-wsd су статистички значајно бољи од S0. Разлика између S7-avg и лексичког приступа L0 има позитиван смер и некоригована вредност теста је испод 0,05, али после корекције остаје на граници (0,0525), па је не треба представити као статистички значајну по примарној макро-F1 мери. Разлика између S7-avg и S7-wsd није статистички значајна, што значи да се у овој поставци не може тврдити да је просек по леми и врсти речи суштински бољи од поступка са синсетом изабраним поступком разрешења вишезначности речи. У целини, низводна евалуација подржава главни закључак да обогачени слој сентимента Српског ворднета има практичну вредност у анализи сентимента реченица. Најјачи доказ односи се на поређење са полазним лексиконом S0: и S7-avg и S7-wsd статистички значајно надмашују директно пресликавање из енглеског SentiWordNet-a. Резултати су повољни и у односу на лексикон L0, али при њиховом тумачењу треба бити опрезан: најбоља конфигурација има више вредности тачности и макро-F1 мере, али разлика у макро-F1 мери после корекције није формално значајна.

5.12 Циљана проверка употребе слоја за разрешење вишезначности речи у машинском превођењу

У овом одељку приказани су резултати циљане провере употребе напомена о значењу у машинском превођењу са српског на енглески језик. Слика 5.18 приказује један пример тога. Циљ провере није био да се оцени општи квалитет преводачких система, већ да се испита да ли експлицитна информација о изабраном значењу може помоћи при избору енглеског преводног еквивалента за лексички вишезначне јединице. Провера је заснована на ручно провереном ELEXIS-WSD-sr материјалу, инвентару значења и енглеском делу ресурса ELEXIS-WSD 1.3. У анализу су укључене само оне српске циљне јединице чије је изабрано значење могло бити повезано са Принстонским ворднетом, јер је поступак формирања напомене захтевао доступне енглеске литерале Принстонског ворднета.



Слика 5.18 Радни пример поређења машинског превода без напомене о значењу и са напоменом о значењу

Основна поставка поредила је превод без додатне информације о изабраном значењу са преводима у којима су у српску реченицу уметнути кратки енглески литерали повезани са изабраним синсетом. У првој варијанти модел је добијао само српску реченицу. У другој варијанти уз сваку доступну циљну јединицу додаван је један енглески литерал из Принстонског ворднета, док је у трећој варијанти додавано до три литерала. У даљем тексту ове поставке се означавају као: основни превод, превод са једним литералом и превод са највише три литерала. Напомене о значењу су намерно биле сведене на минималне преводне трагове: нису садржале енглеску референтну реченицу, ELEXIS дефиниције, глосе,

идентификаторе синсета, нити српске дефиниције. Обим припремљеног скупа, разврстан према највећем броју значења кандидата у реченици, приказан је у Табели 5.35.

Табела 5.35 Обим припремљеног скупа за Google Basic проверу

Опсег броја кандидата значења	Припремљени примери	Циљне јединице	Доступне напомене о значењу	Преводни сегменти
20 и више	16	17	75	48
15-19	80	81	382	240
10-14	474	553	2 194	1 422
Укупно	570	651	2 651	1 710

Поступак се може илустровати реченицом „Брзо лете и тешко их је ухватити.” Глагол *ухватити* има 15 кандидатских значења у инвентару, па је реченица укључена у групу 15–19. Изабраном синсету овог глагола одговарају енглески лтерали *get*, *catch* и *capture*. Напомена је додата и уз облик *лете*, за чије су изабрано значење доступни лтерали *fly* и *wing*. Припремљене су следеће три улазне варијанте за превођење:

Улаз за основни превод

Брзо лете и тешко их је ухватити.

Улаз за превод са једним лтералом

Брзо лете [eng: fly] *и тешко их је ухватити* [eng: get].

Улаз за превод са највише три лтерала

Брзо лете [eng: fly, wing] *и тешко их је ухватити* [eng: get, catch, capture].

У бројеве приказане у Табели 5.35 овај пример улази као један припремљени запис, са једном циљном јединицом из задатог опсега (*ухватити*), двома јединицама са напоменама (*лете* и *ухватити*) и трима улазним варијантама за превођење. Ограничење броја лтерала важи за сваку јединицу посебно: у варијанти са највише три лтерала уз *лете* наведена су два, а уз *ухватити* три лтерала. Лтерали унутар сваке напомене припадају истом изабраном синсету.

Исти пример илуструје проблем задржавања техничких ознака у преводу, који је описан у наставку. Добијени су следећи излази:

Основни превод

They fly fast and are difficult to catch.

Превод са највише три лтерала

They fly fast [eng: fly, wing] *and are difficult to catch* [eng: get, catch, capture].

Глагол *ухватити* у оба случаја преведен је као *catch*, али су у преводу улаза са напоменама задржане и техничке ознаке.

Први увид у резултате показао је да је ова поставка имала практично ограничење: уметнуте напомене су често третиране као видљив део изворног текста, а не као тиха смерница за избор преводног еквивалента. Као што је приказано у Табели 5.36, у *Google Basic* излазима ознаке типа [eng: ...] копиране су у 994 од укупно 1.710 преводних сегмената. Овај налаз показује да сама доступност исправне информације о изабраном значењу није довољна; начин на који се та информација уводи у преводилачки систем може произвести додатни шум.

Табела 5.36 Копирање уметнутих [eng: ...] ознака у *Google Basic* излазима

Највећи број значења кандидата у реченици	Преводни сегменти	Сегменти са копираним [eng: ...] ознакама
20 и више	48	26
15-19	240	136
10-14	1 422	832
Укупно	1 710	994

Због честог копирања ознака и чињенице да је велики број основних превода већ био прихватљив, из пуног скупа је издвојен мањи подскуп за циљану анализу, у ком су примери бирани према променама које су напомене изазвале у преводу. Циљ није био да се формира непристрасан репрезентативни узорак целог корпуса, већ да се пронађу случајеви у којима је напомена о значењу заиста утицала на превод у близини двосмислене циљне јединице. Пре поређења су из превода уклоњене копиране [eng: ...] ознаке, а пример је задржан само ако се очишћени превод са напоменом о значењу довољно разликовао од основног превода и ако су промењени енглески токени били повезани са литералима из Принстонског ворднета за неку циљну јединицу. На тај начин издвојено је 78 примера из првобитних 570 припремљених реченица; њихова расподела према изворном опсегу приказана је у Табели 5.37.

Табела 5.37 Издавање подскупа примера захтевнијих за решавање

Изворни опсег	Изабрани примери
20 и више	2
15-19	11
10-14	65
Укупно	78

Ни у овом подскупу примера захтевнијих за решавање резултати нису показали стабилан образац поузданог побољшања. У бројним случајевима основни превод је већ био прихватљив, док је напомена доводила само до стилске промене или парафразе. На пример, код леме *дати* са литералом *give* (срп. „дати“), основни превод је већ преносио очекивано значење, док је верзија са напоменом само променила *suggest* (срп. „предложити“) у *propose* (срп. „предложити“) и уметнула *give feedback* (срп. „дати повратну информацију“). Та промена није представљала јасну корекцију значења, већ површинско преформулисање. Слично томе, код

леме *напустити*, где су литевали укључивали *abandon* (срп. „напустити“), *forsake* (срп. „оставити, напустити“), *desolate* (срп. „опустошити“) и *desert* (срп. „напустити“), основни превод је већ користио идиоматски одговарајући облик *forsaken* (срп. „остављен, напуштен“), па напомена није довела до јасног побољшања.

Додатна локална провера обухватила је поставке OPUS/Marian и M2M100. Код излаза који је произвео OPUS/Marian уочено је исто основно ограничење: енглеске напомене о значењу често су третиране као део изворне реченице, док прелиминарни преглед излаза није показао довољно стабилне разлике које би оправдале засебну експертску процену. M2M100 је био осетљивији на уметнуте напомене, али та осетљивост није представљала чист доказ побољшања. Дијагностички показатељи за M2M100 приказани су у Табели 5.38: на подскупу од 78 примера захтевнијих за решавање добијена су 234 преводна реда, при чему се копирање ознака јавило у 69 од 78 примера у поставци са једним литевалом и у 76 од 78 примера у поставци са највише три литевала.

Табела 5.38 Дијагностички показатељи за M2M100 на подскупу примера захтевнијих за решавање

Показатељ	Вредност
Број примера	78
Преводни редови	234
Излази са једним литевалом у којима су копиране ознаке	69 / 78
Излази са највише три литевала у којима су копиране ознаке	76 / 78
Медијана минималне сличности између основног превода и превода са напоменом о значењу	0,841
Просечна минимална сличност између основног превода и превода са напоменом о значењу	0,825
Примери испод 0,80 сличности	26 / 78
Примери испод 0,70 сличности	9 / 78
Хеуристички показатељ добитка у избору циљног преводног трага	12 / 78
Хеуристички показатељ губитка у избору циљног преводног трага	1 / 78

Ови резултати показују да уметнуте напомене могу да помере излаз модела, али то померање није било довољно стабилно нити довољно чисто⁵⁴ да би се могло тумачити као поуздано побољшање избора значења. Нижи степен сличности између основног превода и превода са напоменом о значењу често је одражавао копирање ознака, синтаксичко нарушавање реченице или опште погоршање, а не бољи избор преводног еквивалента. Постоје појединачни

⁵⁴ У овом контексту "чисто" се односи на то да су разлике у преводу обухватале и копирање ознака и преформулисање текста, па се сама промена није могла сматрати побољшањем.

позитивни случајеви, на пример, када је напомена померила превод ка *given* (срп. „дат, дато“) уместо погрешног тумачења облика повезаног са лемом *дати*, или ка *voices* (срп. „гласови“) уместо *votes* (срп. „гласови на гласању“). Међутим, такви случајеви су ретки и често праћени другим грешкама у реченици, па не могу бити основ за закључак о практичном побољшању превођења.

На основу ових налаза експеримент је задржан на нивоу прелиминарне евалуације. Добијени излази нису показали довољно јасан и стабилан образац разлика између основног превода и превода са напоменом о значењу да би било методолошки оправдано наставити ка детаљној експертској процени. Таква процена имала би смисла ако би прелиминарни резултати указивали на већи број спорних случајева у којима је потребна фина људска одлука. У овој поставци, међутим, доминирали су случајеви у којима је основни превод већ био прихватљив, у којима је напомена доводила само до парафразирања које нити побољшава нити квари превод, или у којима је уносила шум у превод.

6 Дискусија

У овом поглављу резултати представљени у претходном поглављу тумаче се у односу на постављене хипотезе, релевантна полазишта из прегледа литературе и методолошка ограничења развоја обогашене верзије Српског ворднета. Дискусија је организована као повезано разматрање шест тематских целина: слоја сентимента, разрешења вишезначности речи и редакције инвентара значења, дестилације модела, превода синсета, лексичко-семантичког поравнања ресурса и провере употребе тако обогашених ресурса у низводним задацима. У поглављу се разматра шта добијени резултати показују о границама директног пресликавања постојећих ресурса (преношења вредности сентимента из енглеског SentiWordNet-а на одговарајуће синсете Српског ворднета), о улози великих језичких модела у контролисаном полуаутоматском развоју ресурса и о местима на којима је и даље неопходна експертска лексикографска провера. На тај начин дискусија повезује налазе појединачних експеримената са ширим питањем како се за српски језик може развијати практично употребљив лексичко-семантички ресурс који је подложен провери и проширивању.

6.1 Лексикони сентимента

Резултате добијене при изградњи и евалуацији лексикона сентимента треба тумачити у односу на два полазишта из прегледа литературе. Прво, SentiWordNet је настао као покушај да се ознака сентимента не припише само површинском облику речи, већ појединачним значењима представљеним синсетима. Друго, пренос таквих вредности из енглеског ресурса у друге језике технички је могућ када постоји међујезичко поравнање, али та могућност не подразумева да су добијене вредности довољно поуздане за конкретан језик и конкретан задатак.

Због тога S0 у овој дискусији не треба посматрати произвољно изабрану полазну поставку или као намерно изабран слаби ресурс. Као што је описано у одељцима 2.5 и 2.6, пренос вредности из SentiWordNet-а у друге језике уобичајена је полазна стратегија. Ако приступ који је у другим језичким поставкама употребљив као први корак у Српском ворднету даје ограничене резултате, онда се ограничење не може свести на лош избор полазне поставке. То показује да међујезичко поравнање и еквиваленција превода нису довољни да се очувају поларитет и јачина сентимента конкретног српског синсета.

Ограничење директног пресликавања видљиво је већ у припреми скупова за обучавање. Поједини синсети који су у изворном ресурсу имали позитивне или негативне вредности нису носили одговарајућу вредносну оријентацију у српском језику. Тако су „оценити, оквалификовати, прогласити, изрећи суд о некоме“ и „инфективни агенс“ били спорни као кандидати за позитиван сентимент: први означава чин вредновања, али не подразумева позитиван став: овде је поента да сам чин оцењивања не носи сентимент (за разлику рецимо од красти), док је други неутралан научни појам. Насупрот томе, „Достојевски“ и „удобан“ били су спорни у негативној групи: властито име нема инхерентно негативну оријентацију, а „удобан“ у српском има позитивну конотацију. Вредност сентимента није увек стабилно преносива преко језика, јер зависи од лексичке реализације, грануларности значења, контекста употребе и начина на који је дато значење стварно представљено у српском ресурсу. Зато се директно пресликавање може користити као почетна основа, али не може заменити поступак адаптације и провере.

Евалуације спроведене у овом раду испитују различите аспекте тог проблема. Ручна провера најизраженијих вредности поларитета, синтетички скуп за евалуацију и екстерна анализа

сентимента реченица не дају три неповезана ранга лексикона. Оне представљају три различита филтера истог ресурса. Прва провера испитује да ли лексикон уме да издвоји најснажније позитивне и негативне синсете. Друга провера испитује да ли лексикон уме да очува степенувану структуру поларитета на нивоу синсета, укључујући нулте, слабе, умерене и јаче вредности. Трећа провера испитује да ли се добијене вредности могу употребити у реченицама, где на коначни резултат утичу покривеност, лематизација, морфолошка варијантност, вишезначност, избор синсета и праг за доношење одлуке.

У ручној провери најизраженијих вредности поларитета у први план излазе S4 и S6. Ови лексикони издвајају већи број снажно позитивних или снажно негативних синсета од S0 и зато су информативни када је циљ да се пронађу јасно поларизована значења. Тај налаз, међутим, не значи да су S4 и S6 нужно најбољи кандидати за општи слој сентимента. Лексикон који чешће достиже високе вредности поларитета може бити користан за издвајање екстремних случајева, али истовремено може лакше производити спорне јаке вредности код синсета чија оријентација сентимента није недвосмислена. Управо зато ручна провера најјачих вредности не може бити једини критеријум за избор коначног лексикона.

У ручној провери најизраженијих поларитетских вредности примери треба да остану у дискусији, јер показују природу грешака, а не само бројеве из табеле. Лексикон S1 издвојио је само један снажно позитиван синсет, „фантастичан“, док је у негативној категорији издвојио 26 кандидата. Међу негативним кандидатима били су и спорни случајеви као што су „патак“, „пиједестал“ и „низина“, који не носе изразиту негативну конотацију⁵⁵. S2 је у односу на S1 дао знатно бољу покривеност, са 23 снажно позитивна и 65 снажно негативних синсета, али су и у њему остали спорни примери: „кокетирање, флерт, каћиперство“ у позитивној категорији, а „убеђење, уверење“ и „алтруизам, несебичност, великодушност“ у негативној категорији.

Најуспешнији у издвајању снажно поларизованих синсета био је S4, који је издвојио 55 снажно позитивних и 106 снажно негативних синсета. Ипак, висока покривеност не уклања потребу за ручном провером. У позитивној категорији код S4 спорни су примери „разблудност, похота“, „осигурати, купити“ и „горка чоколада, чоколада за кување“, док се у негативној категорији јављају спорни случајеви као што су „зарадити“, „опустити, релаксирати“, „одржати се, стајати“ и „олакшати, ублажити, утажити“. Сличан образац видљив је и код S6: овај лексикон је издвојио 25 снажно позитивних и 96 снажно негативних синсета, али су у позитивној категорији спорни „подложен, пријемчив“, „осигурати, купити“ и „сетити се, поменути“, а у негативној „дар, поклон“, „разумевање“ и „брижност“.

Насупрот томе, полазни S0, заснован на директном пресликавању из SentiWordNet-a, у ручној провери издваја само шест снажно позитивних и један снажно негативан синсет, „непоштовање“. У позитивној категорији спорни су „истраживач“ и „истраживачки“, док су поједини синсети који се у новим лексиконима јављају као негативни, као што су „тужно“, „себичност“ и „снуждено“, у S0 остали објективни. Ови примери зато показују оба ризика: аутоматски поступци могу произвести лажне јако позитивне или негативне поларитете, док директно пресликавање из SWN може пропустити вредносно обележена српска значења.

Синтетички скуп за евалуацију проверава други аспект квалитета лексикона. На њему најбоље резултате постижу S5 и S7. Лексикон S5, заснован на моделу BERTić, има најмању средњу апсолутну грешку, највишу Спирманову корелацију и најбољу квадратно пондерисану к меру, док S7, заснован на моделу Jerteh-355, постиже веома близак резултат. Овај налаз је значајан јер синтетички скуп није усмерен само на најизраженије поларитете, већ на степенувану

расподелу вредности. Другим речима, S5 и S7 боље прате целу скалу поларитета, а не само њене крајње тачке.

Положај S0 у овој евалуацији посебно је важан за тумачење резултата. Пошто је синтетички скуп за евалуацију с намером формиран из случајева који су у полазном пресликавању из SWN имали вредност (0,0), S0 у овом пресеку задржава нулте вредности за све синсете. Зато његово релативно високо тачно поклапање не треба тумачити као доказ да је директно пресликавање довољно. Оно пре показује да је део кандидата, иако је садржао лексички сигнал из *Senti-pol-sr*, и после провере остао оправдано објективан. То је важан налаз, јер потврђује да присуство поларне лексеме није само по себи доказ поларитета синсета.

Ипак, исти резултат открива суштинско ограничење S0. Пошто су његове вредности у овом скупу константне, S0 не може да разликује оправдано објективне синсете од оних код којих је нулта вредност последица пропуштеног поларитета. Због тога Спирманова корелација за S0 није дефинисана, а квадратно пондерисана к мера једнака је нули. Ове мере показују оно што тачно поклапање само по себи прикрива: директно пресликавање не успоставља употребљиво ординално рангирање синсета према јачини поларитета. У односу на то, предност S5 и S7 није само у мањој просечној грешци, већ у способности да се у српском скупу синсета очува степеност структура вредности сентимента.

Овај резултат непосредно се односи на хипотезу H1. Потврда H1 не почива на једном изолованом налазу, већ на слагању више евалуационих нивоа. Ручно филтрирање скупова за обучавање показује да непосредно пресликавање из SentiWordNet-а производи семантички спорне случајеве у српском ресурсу. Ручна провера најизраженијих поларитета показује да S0 има малу покривеност јаким позитивних и негативних вредности: издваја само шест снажно позитивних и један снажно негативан синсет, а као објективне оставља примере као што су „тужно”, „себичност” и „снуждено”. Синтетички скуп за евалуацију даје јачу унутрашњу потврду: S5 и S7 боље прате степеност вредности поларитета у том скупу, док S0 у овом пресеку остаје на константним нултим вредностима и не успоставља корисно рангирање поларитета. Најзад, екстерна евалуација у анализи реченица даје формалну статистичку потврду у односу на S0, јер S7-avg и S7-wsd статистички значајно надмашују полазно пресликавање после корекције за више поређења.

Због тога H1 треба сматрати подржаном, али уз јасно разграничење снаге доказа. Синтетичка евалуација снажно подржава H1 на нивоу семантике синсета: лексикони добијени поступцима прилагођеним српском ресурсу боље разликују објективне, слабо поларизоване и јаче поларизоване синсете од непосредног пресликавања из SentiWordNet-а. Формално најјачи статистички доказ односи се на поређење S7 са S0 у низводном задатку анализе сентимента реченица. Таква формулација избегава прејаку тврдњу да је сваки аспект H1 потврђен истом врстом теста, али јасно показује да се резултати из методологије, синтетичке евалуације и екстерног задатка међусобно подупиру.

Такво тумачење боље одговара функцији хипотеза у дисертацији. У односу на H1, резултати указују на то да прилагођавање вредности сентимента српском ресурсу може донети мерљиву предност у односу на непосредно преузимање вредности из енглеског SentiWordNet-а. Обим те предности зависи од лексикона и начина евалуације. S5 и S7 показали су најбоље слагање са степеностим референтним оценама у синтетичком евалуационом скупу, док су се S4 и S6 издвојили у ручној провери синсета са најизраженијим поларитетом.

Још један налаз из синтетичке евалуације заслужује посебно тумачење. Сви кандидатски лексикони имају негативно просечно одступање са предзнаком, што значи да у просеку

додељују ниже вредности поларитета од вредности у синтетичком евалуационом скупу. Ово одступање не треба тумачити само као грешку. Оно указује на конзервативан профил добијених лексикона. Такав профил може бити користан ако је приоритет да се избегну лажно поларизовани синсети, али може ограничити осетљивост у задацима у којима је потребно препознати слабе и умерене вредности сентимента. Због тога избор прага у низводној анализи сентимента није само техничко подешавање, већ део начина на који се уравнотежују поузданост и осетљивост ресурса.

Екстерна евалуација у анализи сентимента реченица допуњује проверу вредности сентимента на нивоу појединачних синсета, јер испитује њихову употребљивост за одређивање поларитета целе реченице. У овом задатку резултат зависи од више чинилаца: покривености речи ресурсом, правилног препознавања леме, избора одговарајућег синсета, доступности његових вредности сентимента и прага за одређивање поларитета реченице. Зато ова евалуација показује како вредности сентимента функционишу у оквиру целокупног поступка обраде текста. Она не проверава само да ли су вредности POS(s) и NEG(s) добро распоређене унутар скупа за евалуацију, већ да ли се могу применити у стварним реченицама. У том задатку на резултат утичу и фактори који нису део саме оцене синсета: да ли је релевантна реч покривена ресурсом, да ли је лема правилно препозната, да ли је изабран одговарајући синсет, да ли синсет има доступну вредност сентимента и како је постављен праг за одлуку о поларитету реченице.

У овој евалуацији најјачу потврду добија S7. Конфигурација S7-avg има најбољу макро-F1 меру, док S7-wsd има врло близак резултат и нешто вишу тачност, али нижу покривеност. Обе S7 конфигурације статистички значајно надмашују S0. Тај налаз је посебно важан јер показује да предност обогаћеног слоја синсета није ограничена на унутрашњу евалуацију синсета. Она се преноси и на задатак анализе сентимента реченица у испитаној поставци.

Истовремено, однос S7-avg и S7-wsd показује да прецизнији избор синсета не доноси аутоматски бољи резултат ако је покривеност мања. Просек по леми и врсти речи није семантички прецизнији од варијанте са разрешењем вишезначности речи, али покрива више релевантних јединица. Приступ са разрешењем вишезначности речи ближи је теоријском циљу моделовања сентимента синсета, али у садашњем стању зависи од покривености инвентара значења и доступности вредности сентимента за изабране синсете. Овај резултат није слабост саме идеје о сентименту на нивоу значења, већ показатељ да ресурс, слој за разрешење вишезначности речи и низводни задатак морају да се развијају заједно.

Екстерна евалуација омогућава да се испита да ли се предност обогаћених вредности сентимента, уочена на нивоу синсета, преноси и на одређивање поларитета реченица. У испитаној поставци S7-avg и S7-wsd постигли су боље резултате од S0, а спроведени тестови показали су статистички значајне разлике. Ови налази указују на практичну употребљивост прилагођеног слоја сентимента, уз ограничење да су прагови изабрани на истом аотираном подскупу на коме су израчунати приказани резултати.

Из ових резултата произлази да избор лексикона зависи од планиране употребе, јер примењени поступци евалуације испитују различите аспекте његовог квалитета. S4 и S6 издвојили су се у ручној провери синсета са најизраженијим поларитетом. S5 и S7 показали су најбоље слагање са степенованим референтним вредностима у синтетичком евалуационом скупу, што указује на њихову погодност за представљање различитих степена поларитета у Српском ворднету. У анализи сентимента на скупу од 2.536 реченица из новинских текстова са ручно обележеним

сентиментом, најбоље резултате међу испитаним приступима постигле су варијанте S7-avg и S7-wsd.

Указаћемо и на ограничења ових налаза у наставку дискусије. Синтетички скуп за евалуацију омогућава контролисано поређење лексикона, али није потпуно независан златни стандард. Његова вредност је у томе што се фокусира управо на оне случајеве у којима је полазно пресликавање из SentiWordNet-а било неутрално, а српски лексички сигнал је указивао на могућу поларитетску обележеност. Ручна провера овог скупа показује да је поступак формирања синтетичког евалуационог скупа на нивоу синсета изводљив, али и да спорни случајеви не представљају увек једноставно неслагање са вредношћу коју даје модел: у случајевима у којима није постојало јасно поклапање са предложеном вредношћу није добијена ни стабилна сагласност свих анотатора о истој корекцији. Због тога би у даљем раду било корисно формирати већи, независнији и стратификовано⁵⁶ насумичан евалуациони скуп из целог Српског ворднета, који не би био ограничен само на случајеве неутралног полазног пресликавања и лексичког сигнала из *Senti-pol-sr*. Такав скуп би омогућио поузданију проверу да ли се добијени закључци преносе на објективне, слабо, умерено и снажно поларизоване синсете у ширем ресурсу. Екстерна евалуација, с друге стране, показује практичну употребљивост ресурса, али је заснована на реченицама преузетим из новинских текстова. Због тога резултате не треба проширивати на све домене српског језика без додатне провере.

Ови налази усмеравају даљи рад у два правца. Први правац је лексикографски и односи се на ресурсе: потребно је ширити покривеност синсета вредностима сентимента, ускладити слој сентимента са инвентаром значења и систематски преиспитивати случајеве у којима ознака поларитета лексема у лексикону Senti-pol-sr не одговара њеном конкретном значењу у синсету. Други се односи на модел: могуће је испитати поступке у којима би фино подешен трансформерски класификатор директно давао калибрисане вероватноће за POS, NEG и OBJ, или независне вероватноће за позитивну и негативну компоненту. Такав приступ би био ближи савременим моделима класификације, али би морао да се провери истим критеријумима као и постојећи лексикони: евалуацијом синсета, ручном провером, екстерним задатком, контролом доменског преноса и провером калибрације вероватноћа.

У целини, резултати показују да директно пресликавање из SentiWordNet-а остаје корисна полазна основа, али није довољна за слој сентимента Српског ворднета. Најбољи резултати не произлазе из механичког преузимања вредности из енглеског ресурса, већ из поступака који комбинују постојеће ресурсе, машинско учење, проверу на текстовима на српском и евалуацију на више нивоа. Главни закључак који се тиче увођења слоја сентимента у Српски ворднет јесте да вредности преузете из енглеског SentiWordNet-а треба проверити и, по потреби, прилагодити конкретним значењима српских лексема у одговарајућим синсетима. Употребљивост тако добијених вредности потребно је проверити и на нивоу појединачних синсета и у задацима обраде текстова на српском језику.

6.2 Разрешење вишезначности речи и редакција инвентара значења

Резултати разрешења вишезначности речи непосредно се надовезују на полазни аргумент дисертације: велики језички модели могу да допринесу аутоматском разумевању текста, али

⁵⁶ Стратификован случајни избор значи да се примери насумично бирају унутар унапред издвојених група; број примера по групама одређује се посебно евалуациони скуп формиран случајним избором синсета унутар унапред одређених група према категорији и степену поларитета-

њихова употребна вредност у раду са Српским ворднетом зависи од експлицитно моделованог лексичко-семантичког знања. У уводу је овај однос формулисан кроз општу хипотезу *H0*, која повезује аутоматизацију развоја семантичких мрежа са поузданим обележавањем текста синсетима, као и кроз хипотезу *H2*, у којој се испитује могућност придруживања циљне лексеме одговарајућем синсету. Зато резултате из дела рада посвећеног ELEXIS-WSD-sr корпусу не треба тумачити само као поређење модела, већ као проверу шире методолошке идеје: да се корпусни примери, инвентар значења, велики језички модели и ручна провера могу повезати у контролисан поступак развоја ресурса.

Овај налаз има јасну везу са прегледом литературе у поглављу 2. Разрешење вишезначности речи је у литератури описано као један од основних задатака рачунарске семантике, јер се односи на избор конкретног значења полисемичне лексеме у датом контексту (Navigli 2009). Рани приступи ослањали су се на глосе, речничке дефиниције и структуру лексичко-семантичких ресурса, док савремени приступи све више користе контекстуалне репрезентације и велике језичке моделе. Међутим, резултати овог рада потврђују да се ова два правца међусобно не искључују. Напротив, велики језички модели су најкориснији онда када њихов излаз није слободно парафразирање значења, већ контролисан избор унутар унапред дефинисаног инвентара. На тај начин се савремена способност модела да тумачи контекст повезује са стабилношћу ворднета, односно са синсетима, дефиницијама и идентификаторима који омогућавају проверљиву анотацију.

Методолошка вредност овакве поставке лежи у томе што модел не делује као самостални лексикограф, већ као део полуаутоматског поступка. За сваку циљну јединицу модел добија реченични контекст, лему, врсту речи и ограничен скуп значења кандидата из инвентара. Ово је важно за српски језик, јер морфолошка сложеност и велики број површинских облика чине да се анотација не може ослонити само на подударане облике речи. Лема и врста речи служе као веза између конкретне појаве речи у тексту и релевантног дела инвентара, док контекст реченице омогућава разликовање конкурентних значења. Таква организација задатка уједно смањује простор за неконтролисано генерисање и омогућава да се излаз модела повеже са стабилним инвентаром синсета.

Изградња почетног инвентара WSD-sr-sens-inv-v1 показује да овај поступак није био једноставна примена готовог ресурса. Српски ворднет је послужио као основа, али није имао потпуну покривеност свих лексема и значења потребних за обележавање корпуса. Зато је изградња почетног инвентара морала да прође кроз више повезаних корака: поравнање са енглеским ELEXIS-WSD-en инвентаром и Принстонским ворднетом, издвајање приоритетних синсета из реченичних анотација, превођење и ручна дорада дефиниција, додавање хиперонима ради очувања хијерархијске повезаности, као и укључивање јединица које нису биле непосредно заступљене у енглеском ресурсу. У резултатима је посебно наглашено да је у овој фази додато 2.150 синсета, а да је Српски ворднет после проширења обухватао 29.629 синсета. Ови подаци су важни за дискусију јер показују да разрешење вишезначности речи није било накнадно примењено на статичан ресурс, већ је захтевало претходно проширивање и прилагођавање самог инвентара значења.

Примери из изградње почетног инвентара додатно показују специфичност српског језичког материјала. За неке јединице било је могуће ослонити се на описне речнике српског језика, нарочито код придева и прилога који немају непосредан енглески еквивалент или нису заступљени у Принстонском ворднету, као што су *дувански* и *зими*. Посебну групу чиниле су изведене именице женског рода, деминутиви и аугментативи, на пример *новинарка* и *јајашце*, као и појмови који нису били заступљени у Принстонском ворднету, као што је *дума*. За

терминолошке јединице, новије лексеме и вишечлане изразе, на пример *хадејк* (најранији геолошки еон у историји Земље, који је претходио архаику) и *мастер-клас* (лекција или курс који води стручњак високог нивоа у одређеној области), дефиниције су обликоване уз помоћ великих језичких модела и потом ручно редиговане. Ови примери су битни јер показују да се развој српског инвентара значења не састоји само од пресликавања енглеског инвентара, већ је то поступак у коме се међујезичко поравнање, локални лексички извори, подршка генеративних модела и људска процена морају комбиновати.

Прелазак са верзије WSD-sr-sens-inv-v1 на WSD-sr-sens-inv-v2 још јасније показује да модел и ресурс делују заједно. Друга верзија није само механички већа од прве, већ је настала на основу аутоматске анотације, ручне провере и анализе случајева у којима постојећи инвентар није обезбеђивао одговарајуће значење. У њу су укључена недостајућа значења која су у претходној фази била обележавана ознаком NEW_SENSE, или је њихов недостатак уочен у развојно-аналитичком подскупу провере од ручно реченица. Поред додавања нових значења, побољшано је 213 дефиниција. Тако је, на пример, дефиниција лексеме *девојчица* прецизирана у односу на претходну широку формулацију, док је дефиниција глагола *учествовати* замењена информативнијом формулацијом која боље описује типичну употребу. Овакве измене показују да редакција није била само лексичка, већ и семантичка и лексикографска. Овакве измене показују да је редакција обухватила и проширивање инвентара недостајућим значењима и прецизирање дефиниција већ уврштених значења.

Поређење две верзије инвентара потврђује исту тезу. Број јединица у инвентару значења порастао је са 12.747 на 13.227, број значења са 9.929 на 10.284, док је број лема незначајно увећан, са 6.108 на 6.122. Овај однос је важнији од самог апсолутног пораста. Он показује да редакција није била усмерена на проширивање списка лема, него на повећање семантичке покривености постојећих. Другим речима, главни добитак није био у томе што је инвентар обухватио велики број нових лема, већ у томе што је боље покрио разлике значења које су се показале релевантним у корпусу. Управо то је битно за разрешење вишезначности речи: модел може да изабере исправно значење само ако је оно присутно у инвентару и ако је довољно јасно разграничено од суседних значења. Другим речима, инвентар је проширен првенствено значењима која су се показала релевантним у корпусу, уз мали пораст броја лема. За успешно разрешење вишезначности речи важно је да одговарајуће значење буде присутно у инвентару и да његова дефиниција јасно показује по чему се оно разликује од других кандидатских значења.

Евалуација у две итерације зато има шири значај од поређења појединачних модела. У првој итерацији коришћен је почетни инвентар, док је у другој коришћена ревидирана верзија. Побољшање резултата у другој итерацији показује да квалитет аутоматске семантичке анотације не зависи само од способности модела, већ и од покривености, грануларности и јасноће инвентара значења. На издвојеном тест скупу од 300 реченица GPT-4.1 у другој итерацији достиже 92,4% тачности за једночлане и 95,2% за вишечлане јединице. Овај резултат представља јаку унутрашњу потврду хипотезе *H2* у делу који се односи на придруживање текстуалних појава одговарајућим синсетима. Истовремено, он не значи да је задатак решен безусловно, јер се добијена тачност ослања на претходно припремљен инвентар, пажљиво обликован промпт и накнадну ручну проверу. Наведена тачност остварена је на издвојеном тест скупу, уз употребу ревидираног инвентара значења и примењеног упита. За процену успешности поступка на другим текстовима, као и у случајевима у којима инвентар не садржи одговарајуће значење, потребна је додатна евалуација.

Посебно је важно раздвојити улоге ручно проверених података. Развојно-аналитички скуп од 120 реченица био је део итеративног развоја поступка: на њему су уочаване грешке, издвајани случајеви непокривених значења и формулисани основи за редакцију инвентара. Насупрот томе, издвојени тест скуп од 300 реченица има већу тежину у процени текуће верзије поступка, јер није коришћен за накнадно прилагођавање модела, промпта или инвентара након прве итерације. Укупно гледано, златни скуп ELEXIS-WSD-sr-gold садржи 420 ручно проверених реченица и 2.693 аотирана сегмента текста, од чега 2.430 једночланих јединица и 263 вишечлане јединице. Он зато није само средство за мерење тачности, већ и опис тежине самог задатка: део примера има само једно кандидатско значење, док су други примери заиста двосмислени и укључују више конкурентних значења.

Ова разлика је важна за тумачење резултата. Случај са једним кандидатом не представља исти тип проблема као случај у коме модел мора да бира између шест, десет или више значења. У првом случају главни изазов је покривеност и стабилно повезивање леме са инвентаром, док је у другом случају изазов стварно разликовање блиских значења у контексту (за она која су семантички удаљена није тешко одредити). Зато просечна тачност, иако неопходна као општа мера, не сме да прикрије разлику између тривијалних, умерено тешких и високо полисемичних примера. За будуће верзије поступка корисно је да се евалуација систематски приказује и према броју кандидатских значења, врсти речи и типу грешке, јер управо такве анализе показују где инвентар треба проширити, а где треба прецизније разграничити постојећа значења.

Квалитативна анализа грешака је због тога један од најважнијих делова резултата. Она показује да грешке не треба тумачити само као неуспех модела, већ као дијагностички материјал за даљи развој ресурса. Најизраженији проблеми јављају се код лема са развијеном полисемијом, код глагола чије значење зависи од аргументске структуре и ширег синтаксичког окружења, као и код случајева у којима су дефиниције блиске или недовољно разграничене. У анализи су као проблематичне издвојене леме као што су: *земља*, *одређен*, *добити*, *број* и *моћи*. Такви примери показују да није довољно додати још једну дефиницију у инвентар; често је потребно додати типичан контекст, контрастивне примере или јасније описати границу између суседних значења.

Посебно су значајни глаголи. Код њих избор значења зависи од аргументске структуре, конструкције и ширег реченичног оквира, па сама кратка дефиниција не мора бити довољна за поуздано разликовање. Ако модел меша блиска глаголска значења, проблем може бити у моделу, али може бити и у томе што инвентар не садржи довољно јасне показатеље употребе. У том смислу, грешке код глагола треба посматрати као приоритетну листу за будућу редакцију: потребно је додавати типичне аргументе, конструкције, примере употребе и, где је могуће, контрастивне напомене које разликују блиске синсете. Таква редакција није споредни технички посао, већ услов да се инвентар стварно примени у контекстуалним задацима.

Ознака NEW_SENSE има посебан методолошки значај. Она се на први поглед може посматрати као неуспешан исход, али је за развој ресурса управо таква ознака један од најкориснијих сигнала. Када модел не може да изабере ниједно од понуђених значења, то може указивати на недовољну покривеност инвентара, на нејасну дефиницију, на превише широко или превише уско постављено значење, или на проблем у филтрирању кандидата синсета. Због тога поставка која дозвољава NEW_SENSE не мора увек да максимизује формалну тачност у процесу евалуације, али омогућава да се аутоматска аотација претвори у механизам за откривање празнина у ресурсу. У томе је њена вредност за дисертацију: она раздваја поступак који бира најближи од понуђених кандидата од поступка који може да укаже да понуђени кандидати нису довољни.

Овај закључак је у складу и са резултатима аблационих⁵⁷ провера промпта. Пуни промпт, који укључује образложење и могућност избора NEW_SENSE, има другачију улогу од скраћеног промпта, који модел усмерава само ка избору једног од понуђених кандидата. Уклањање NEW_SENSE у неким поставкама може повећати формалну тачност, јер модел више нема могућност да се уздржи од избора. Међутим, такво повећање није нужно боље са становишта редакције ресурса, јер се губи информација о случајевима у којима инвентар није довољан. Зато најбоља поставка за развој инвентара није увек она која даје највећи проценат тачних одговора у затвореном избору, већ она која истовремено даје употребљиву анотацију и сигнале за даљу лексикографску проверу.

Употребу великих језичких модела у овом поступку зато не треба тумачити као замену за људску анотацију, већ као промену распореда анотационог рада. Модел може брзо да предложи значење, рангира кандидате, издвоји случајеве у којима постоји сумња на непокривено значење и понуди образложење за избор. Ипак, тиме не нестаје потреба за експертском провером. Људски рад се помера са почетног означавања сваког појединачног примера на контролу излаза модела, анализу спорних случајева, допуну дефиниција и редакцију граница између блиских значења. То је суштина полуаутоматског поступка: аутоматизација убрзава и усмерава рад, али не уклања одговорност за коначну лексикографску одлуку (Petalinkar et al. 2026).

Отвара се и питање цене поступка која се не ограничава само на новчани трошак коришћења појединачног модела. Цена обухвата рачунарски трошак или трошак API позива код јачих модела, време људске провере, цену погрешне анотације, потребу за локалним извршавањем и вредност информације која се добија из грешака. Ако је циљ највећа могућа тачност на ограниченом и важном скупу примера, резултати оправдавају употребу јачих модела уз ручну проверу. Ако је циљ обрада већег корпуса, онда локални или дестиловани модели могу бити прихватљивији као први пролаз, под условом да се њихови излази не третирају као коначна анотација без додатне контроле. Према томе, у пракси питање није само који модел је најтачнији, већ који распоред рада најбоље одговара расположивим ресурсима, прихватљивом ризику и намени анотације.

Резултати дестилације показују у којој мери се анотације из полуаутоматског поступка могу искористити за обучавање мањих модела за разрешење вишезначности речи. Најбољи дестиловани модел, *multilingual-e5-large*, достиже 72,6% тачности на издвојеном тест скупу, док *Llama4* достиже 76,9%, а *GPT-4.1* 92,9% у упоредивом приказу. Иако заостају за најуспешнијим генеративним моделом, дестиловани модели могли би имати практичну улогу у обради већих количина текста: у припремној анотацији, рангирању кандидатских значења и издвајању примера за експертску проверу. Корисност ових примена потребно је засебно испитати.

Овај налаз има и шири значај за преносивост поступка. Ако се слична методологија примени на други језик, други ворднет или други корпус, није довољно само заменити језик у промпту. Потребно је проценити квалитет инвентара значења, начин издвајања кандидата односно одговарајућих синсета, доступност лема и морфосинтаксичких ознака, однос између локалног ворднета и Принстонског ворднета, као и доступност ручно провереног скупа за евалуацију. За језике са мање развијеним ресурсима посебно је важно да поступак предвиди механизам за

⁵⁷ Аблациона проверка подразумева изостављање појединих компоненти поступка и поређење резултата са потпуном верзијом, како би се проценио допринос тих компоненти. Овде се односи на поређење пуног упита са варијантама из којих су уклоњени захтев за образложењем или могућност избора ознаке NEW_SENSE.

решавање проблема непокривених значења, јер би принудни избор из непотпуног инвентара могао дати формално уредне, али семантички погрешне анотације.

У односу на хипотезе, резултати разрешења вишезначности речи најдиректније подржавају $H2$ и онај део $H0$ који се односи на поуздано обележавање текста синсетима и унапређење развоја семантичке мреже. Они делимично и у ограниченом смислу поткрепљују и $H3$: велики језички модели су коришћени за контролисану допуну дефиниција и подршку редакцији, док су њихови предлози пролазили кроз ручно редиговање и проверу. С друге стране, ови резултати сами по себи не доказују успешност низводних задатака. За анализу сентимента и машинско превођење потребно је засебно испитати како се информација о значењу интегрише у конкретан систем.

У целини, резултати представљени у одељку о разрешењу вишезначности речи показују да се за српски језик може изградити ручно проверен евалуациони материјал и да се ВЈМ могу успешно применити у режиму контролисаног избора значења. Још важније, резултати показују да се исти поступак може употребити као механизам развоја ресурса: он открива недостајућа значења, нејасне дефиниције, преклапања у инвентару и типове лексема које захтевају додатну обраду. Због тога је оправдана примена аутоматизације у итеративном поступку разрешења вишезначности речи који се ослања на јасно дефинисан инвентар значења и укључује ручну проверу резултата.

6.3 Дестилација модела и ограничења мањих модела

Резултати дестилације непосредно се настављају на налазе из претходног одељка, али имају другачији фокус. У одељку 6.2 разрешење вишезначности речи разматра се као механизам изградње и редакције инвентара значења: грешке и случајеви недодељеног значења (NEW_SENSE) показују где инвентар треба проширити или прецизније разграничити поједина значења. У овом одељку питање је уже и оперативније: да ли се знање настало у том полуаутоматском поступку може пренети на мањи модел који је погоднији за бржу, локалну и јефтинију примену над већом количином текста. Због тога дестилацију не треба тумачити као засебан покушај да се замени поступак представљен у претходном одељку, већ као његов оперативни наставак.

Такво постављање проблема у складу је са прегледом литературе о генерисању података, сребрним скуповима и дестилацији модела. У условима ограничених језичких ресурса, велики језички модели могу да произведу већи број аутоматски означених примера, али ти примери немају исти статус као ручно проверени подаци. У овој дисертацији зато је направљена јасна разлика између две улоге података. Сребрни скуп служи као материјал за обучавање мањег модела: он преноси одлуке добијене у полуаутоматској анотацији и омогућава да се модел обучи на већем броју примера који се не би могао произвести само ручним означавањем. Златни скуп, насупрот томе, не служи за обуку, већ за финалну евалуацију. Његова вредност је управо у томе што остаје ручно проверена и методолошки одвојена основа за процену да ли је пренос знања био успешан.

Ово разграничење је важно јер спречава погрешно тумачење дестилације. Циљ није да се сребрни скуп прогласи коначном основом за евалуацију, нити да се мањи модел оцени на истом материјалу из ког је учио. Циљ је да се провери да ли се сребрни скуп, упркос могућим грешкама у аутоматски додељеним ознакама значења, може употребити за обуку модела који ће рангирати кандидатска значења у новим примерима. Поузданост тог преноса процењује се

тек на златном скупу. У том смислу, дестилација представља тест употребљивости сребрне анотације: ако мањи модел после обуке постиже боље резултате на ручно провереном тесту, то значи да је у сребрном скупу постојао преносив сигнал, а не само акумулација непроверених ознака.

Методолошки посматрано, задатак мањег модела није постављен као глобална класификација свих значења у инвентару, већ као процена сагласности између контекста у реченици и дефиниције кандидатског значења. За сваку циљну јединицу модел упоређује реченицу са понуђеним дефиницијама које одговарају датој леми и врсти речи. Та формулација задатка важна је из два разлога. Прво, она задржава везу са инвентаром значења, па мањи модел не предвиђа произвољну ознаку, већ бира у оквиру контролисаног скупа кандидата. Друго, она омогућава да се излаз модела тумачи као рангирање значења, што има практичну корист у поступцима где се очекује накнадна провера експерата.

Резултати показују да обука на сребрном скупу може да побољша мање моделе, али да се тај ефекат разликује међу испитаним моделима. Код модела *simple* тачност расте са 54,4% на 68,5%, док код модела *multilingual-e5-large* расте са 59,2% на 72,6%. Оба побољшања су статистички значајна према Макнимаровом тесту уз Холмову корекцију. Код модела *tesla* тачност опада са 53,4% на 51,3%, али тај пад није статистички значајан на нивоу 0,05 (коригована р-вредност = 0,0822). Ови налази показују да је корист од дестилације потребно проверити за сваки модел у конкретној поставци обуке.

Неуспех модела *tesla* у овој поставци посебно је поучан. Пошто је реч о моделу прилагођеном српском и сродном језичком простору, могло би се очекивати да ће језичка блискост сама по себи донети предност. Резултати, међутим, показују да то није довољно. За овај задатак није пресудно само да модел добро представља српски текст, већ да уме да упари конкретну употребу речи у реченици са кратком дефиницијом једног могућег значења. Ово ограничење треба посматрати и у односу на природу претходне обуке модела: *tesla* је обучен кроз предвиђање маскираних токена, чиме стиче опште језичке репрезентације,⁵⁸ али такав циљ обуке није исто што и оптимизација за семантичко поређење контекста и дефиниције. Зато се део језичког знања може пренети на задатак над српским материјалом, али тај пренос не мора бити довољан ако архитектура, простор репрезентација или додатна обука на сребрном скупу не подрже управо овај тип семантичког упаривања.

Додатна анализа показује да добитак од дестилације није последица тривијалних случајева. У исправљеној евалуацији примери са само једним кандидатом значења третирали су одвојено, јер у њима нема стварног избора између више значења. Најважнији добици код модела *simple* и *multilingual-e5-large* јављају се управо код једночланих јединица (речи) и код заиста двосмислених примера, нарочито тамо где постоји и више од 3 кандидатских значења. Код јединица са 11 или више кандидата *simple* расте са 2,1% на 40,0%, а *multilingual-e5-large* са 9,5% на 45,3%, при чему проценти означавају тачност избора значења у групи од 95 тест примера са најмање 11 кандидата, пре и после дестилације. Ови резултати показују да дестилација не побољшава само покривеност или техничко повезивање ознака, већ помаже моделу да боље разликује конкурентна значења у контексту.

⁵⁸ „Опште језичке репрезентације“ означавају нумеричке приказе речи и њиховог контекста који одражавају обрасце научене из текстова. Ово ограничење треба размотрити и у односу на претходну обуку модела *tesla*. Предвиђањем маскираних токена, односно привремено скривених речи или делова речи, модел учи обрасце њихове употребе и повезивања у контексту. Такав циљ обуке није непосредно усмерен на избор дефиниције која одговара значењу речи у конкретној реченици.

Ипак, перформансе дестилованих модела остају испод најуспешнијих генеративних поставки. На издвојеном тест скупу најбољи дестиловани модел, *multilingual-e5-large*, достиже 72,6% тачности, док *Llama4* достиже 76,9%, а *GPT-4.1* 92,9%. Овај однос треба тумачити опрезно. Дестиловани модел није замена за најпрецизнији генеративни поступак када је циљ највећа могућа тачност на ограниченом скупу. Његова вредност је другачија: он омогућава бржу и јефтинију примену над већим бројем примера, уз задржавање везе са инвентаром значења и уз могућност да се резултати користе као рангирани предлози, а не као неконтролисана коначна анотација.

Из овога следи итеративна логика целог поступка. Бољи инвентар значења омогућава поузданију полуаутоматску анотацију; поузданија анотација даје квалитетнији сребрни скуп; већи и квалитетнији сребрни скуп може побољшати обуку мањих модела; а постепено проширивање златног скупа омогућава строжу и стабилнију евалуацију. Међутим, овај ланац није аутоматски. Ако сребрни скуп садржи систематске грешке, ако су у инвентару недовољно разграничена блиска значења или ако мањи модел не уме да представи однос између контекста и дефиниције, повећање количине података не мора довести до бољег резултата. Због тога се дестилација мора посматрати као део итеративног радног тока, а не као завршна фаза аутоматизације.

Практична вредност дестилованих модела најјасније се види у организацији рада са већим корпусом. Мањи модел може да служи као први пролаз кроз материјал, као средство за рангирање значења кандидата и као начин да се издвоје случајеви који заслужују ручну проверу. Посебно вредни могу бити примери у којима се дестиловани модел и генеративни модел не слажу, као и примери у којима је разлика између најбоље ранжираних кандидата мала. Такви случајеви вероватно указују на блиска значења, недовољно информативне дефиниције или контексте који захтевају додатно лексикографско тумачење. Дестилација, дакле, може помоћи да се људски рад не троши равномерно на све примере, већ да се усмери на оне који имају највећу дијагностичку вредност.

Овај закључак има значај и за ширу примену обогаченог Српског ворднета, али га не треба мешати са евалуацијом низводних задатака. У одељку 6.6 разматра се како се обогачени ресурс понаша у анализи сентимента и машинском превођењу. Овде је важнији инфраструктурни аспект: дестилација показује да се подаци настали у поступку разрешења вишезначности речи могу искористити за изградњу помоћног модела који скалира рад са ресурсом⁵⁹. Такав модел у будућности може бити део ширег радног тока за обраду текста, укључујући избор значења, издвајање спорних случајева и припрему материјала за задатке као што су анализа сентимента или семантичко претраживање, али његова примена у тим задацима мора се посебно евалуирати.

Ограничења ових резултата треба јасно нагласити. Прво, златни скуп је ручно проверен, али је и даље ограничен по обиму, па будуће итерације треба да га проширују, нарочито примерима са већим бројем значења кандидата и глаголским значењима. Друго, сребрни скуп има нижу поузданост од златног скупа и може пренети систематске грешке из полуаутоматске анотације. Треће, резултати на једном тест скупу не гарантују исту стабилност на другим корпусима или доменима. Зато будући рад треба да провери да ли се исти однос између сребрног скупа, дестилованог модела и златне евалуације одржава када се поступак примени на нове корпусе и на већи број ручно проверених примера.

⁵⁹ Мисли се на примену инвентара значења при обради већих количина текста, уз помоћ модела који рангира кандидатска значења

У целини, анализа резултата дестилације подржава општи аргумент дисертације у ограниченом, али важном смислу. Резултати не показују да мањи модели могу да достигну квалитет најбољих генеративних поставки, али показују да се знање настало у полуаутоматском поступку може делимично пренети на ефикасније modele. Најјача тврдња зато није да дестилација решава проблем разрешења вишезначности речи, већ да омогућава оперативни слој између скупе генеративне анотације и ручне провере. Управо тај слој је важан за одржив развој Српског ворднета: он повезује сребрне податке, златну евалуацију, инвентар значења и људску ревизију у поступак који се може постепено проширивати.

6.4 Превод синсета помоћу великих језичких модела

Резултати превода синсета помоћу великих језичких модела непосредно се односе на хипотезу *H3*, у којој се испитује да ли се ВЈМ могу употребити за контролисану допуну синсета, укључујући формулисање дефиниција и предлагање синонимних или лексичких варијанти. Овај одељак зато не треба читати само као поређење три техничка радна тока, већ као проверу ужег дела хипотезе *H3*: да ли генеративни модел може да произведе лексикографски лексикографски употребљивог кандидата за српски синсет и да ли је такав поступак стабилнији од непосредног превођења глосе и литерала.

Такво постављање задатка надовезује се на претходна истраживања о аутоматском и полуаутоматском проширивању ворднета. У тим истраживањима показано је да превођење из већег ворднета може убрзати изградњу ресурса, али и да преводни кандидати морају проћи рангирање, семантичко филтрирање и ручну или полуаутоматску валидацију (Lam et al. 2014; Al Tarouti and Kalita 2016). Новији радови са великим језичким моделима померају тежиште са превођења појединачних лема на генерисање ширег лексичко-семантичког садржаја, али и даље наглашавају потребу за евалуацијом и редакторском контролом (Bergh et al. 2025; Santoro et al. 2025). Због тога је очекивано да полазни радни ток, заснован на непосредном преводу, буде најслабији, као и да поступци који са више ослањају на семантички контекст буду погоднији за формирање глоса.

Резултати у целини подржавају тај очекивани смер, али га и прецизирају. Полазни радни ток има најнижи просек од пет критеријума за глосу, што потврђује да непосредно превођење није довољно да би се добио лексикографски употребљив српски запис. Тај налаз је у складу са претходним истраживањима у којима су приступи који се ослањају само на превод корисни као почетна стратегија, али не и као завршно решење. Ово значи да хипотеза *H3* није подржана у јаком облику ако би се тумачила као могућност непосредног аутоматског преноса енглеског синсета у српски, али јесте подржана у делу који се односи на употребу ВЈМ за припрему контролисаног скупа кандидата.

Најубедљивији описни налаз у прилог хипотезе *H3* долази из поређења полазног и концептуалног радног тока. Концептуални ток има највиши просек оцена које се односе на глосе и највише случајева у којима глоса не захтева исправку. То показује да кратка енглеска глоса и листа литерала нису увек довољан улаз за добар српски синсет. Када модел најпре формира проширену репрезентацију значења, а тек затим приступа избору литерала и формулисању српске глосе, добија се природнији запис који то значење одваја од осталих. Овај налаз је у складу са ширим увидима из инжењерства промптова: сложени задаци захтевају јасно формулисан задатак, ограничења, структуриран излаз и, када је потребно, раздвајање анализе од коначне формулације (Reynolds and McDonell 2021; Schulhoff et al. 2024).

Мање очекиван резултат односи се на вишефазни радни ток. На основу литературе о уланчавању промптова и самопобољшавање (енгл. *self-refinement*)⁶⁰ могло би се очекивати да разлагање задатка на више корака стабилизује излаз, јер модел одвојено анализира синсет, предлаже кандидате, филтрира их и формулише глосу (Madaan et al. 2023; Schulhoff et al. 2024). У овом експерименту, међутим, вишефазни ток није био стабилнији од концептуалног. Његове глосе су често добро структуриране и нису циркуларне, али је адекватност понуђених литерала знатно слабија, а сувишни литерали су забележени у највећем броју проверених редова. Ово показује да већи број корака не значи нужно већу лексикографску поузданост ако се у некој фази скуп кандидата проширује без довољно строгог семантичког ограничења.

Овај налаз је важан за тумачење хипотезе *H3*, јер разграничава две врсте контроле. Формална контрола излаза може побољшати структуру глосе, редослед корака и машинску обрадивост резултата, али не гарантује да су предложени литерали чланови истог синсета. Лексикографска контрола је захтевнија: она подразумева проверу лексикализације,⁶¹ морфолошке прикладности, односа према блиским значењима и стварне употребе у српском језику. Зато се квалитет синсета не може процењивати само на основу природност дефиниције или формалне уредности поступка.

У том смислу, резултати показују да хипотезау *H3* треба сматрати делимично подржаном. Велики језички модели могу да предложе глосе које су у духу српског језика и користан почетни скуп литерала, нарочито када је поступак организован тако да се пре коначне формулације стабилизује значење синсета. Међутим, резултати не подржавају јачу тврдњу да се Српски ворднет може поуздано проширивати непосредним аутоматским генерисањем. У сва три радна тока уочено је да има литерала који недостају као и оних који су сувишни, а у појединачним примерима јављају се кандидати који имају блиско али не и тражено значење, који су преузети из енглеског или су неприхватљиво генерисани и не припадају српском језику. Због тога хипотеза *H3* потврђује употребљивост ВЈМ као дела полуаутоматског поступка, али не укида потребу за експертском лексикографском редакцијом.

Овај закључак се надовезује и на општу хипотезу *H0*. Резултати показују да је могуће аутоматизовати део процеса развоја семантичке мреже: модел може припремити глосу, предложити литерале и издвојити материјал који лексикограф затим проверава. Ипак, аутоматизација⁶² није исто што и потпуно самостално уношење нових синсета у ресурс. Најстабилнији модел рада је итеративан: ВЈМ припремају кандидатски запис, а ручна провера одлучује о коначној глоси, прихватљивим литералима и односу према постојећој структури Српског ворднета.

Тумачење ових резултата је ограничено малим обимом ручно провереног подскупа. Један експерт је валидирао 30 синсета, односно 90 редова са кандидатима. Зато резултате треба представити као описну евалуацију која показује тенденције и врсте проблема, а не као коначну статистичку потврду предности једног поступка. Ипак, унутар тог подскупа подаци доследно

⁶⁰ Итеративна дорада сопственог одговора је поступак у коме модел најпре генерише одговор, затим процењује његове недостатке и на основу сопствених примедба израђује измењену верзију. Провера и дорада могу се понављати у више корака. Дорађује се генерисани одговор, без додатне обуке модела.

⁶¹ провера лексикализованости подразумева проверу да ли предложени литерал представља устаљену реч или израз у српском језику са одговарајућим значењем, јер за литерал није довољно да буде разумљив опис неког појма.

⁶² У овом случају аутоматизација обухвата припрему кандидатског записа: ВЈМ предлаже глосу и скуп литерала. Лексикограф затим проверава предложено значење, језичку прихватљивост глосе и литерала и однос према постојећим синсетима, па на основу те провере одлучује о потребним исправкама и могућности уношења записа у Српски ворднет.

указују на две тенденције: полазни ток је најслабији у формулисању глоса, док је вишефазни ток најризицијнији у погледу прекомерног додавања литерала. За чвршћи закључак потребно је проширити проверени подскуп, обухватити све врсте речи у већем броју примера, укључити бар још једног независног експерта у процес валидације истог скупа или барем за циљану проверу спорних случајева, а посебно анализирати случајеве у којима се српски литерали не могу непосредно извести из енглеских литерала.

У целини, овај део истраживања допуњује главну аргументацију дисертације на исти начин као и истраживања о изградњи лексикона сентимента и разрешењу вишезначности речи. Најстабилнији резултат не добија се механичким преносом енглеског ресурса на српски, већ контролисаном комбинацијом аутоматске обраде, експлицитне семантичке репрезентације и ревизије од стране експерата. У односу на хипотезу *H3*, најпрецизнији закључак гласи: велики језички модели јесу употребљиви за контролисану припрему кандидата за глосе и литерале, али је проширивање Српског ворднета методолошки оправдано само као полуаутоматски поступак у коме је експертска провера саставни део поступка, а не накнадни формални додатак.

6.5 Поравнање лексичко-семантичких ресурса

Резултати поравнања описног речника и Српског ворднета непосредно се односе на онај део хипотезе *H2* који се тиче поравнања значења на нивоу речничких дефиниција и синсета, али се истовремено надовезују и на општи оквир хипотезе *H0*, у којој се испитује могућност да се методама машинског учења и великим језичким моделима аутоматизују и унапреде процеси развоја ворднета као ресурса. У претходном делу дискусије хипотеза *H2* је разматрана у односу на разрешење вишезначности речи у тексту, где модел треба да изабере одговарајући синсет за конкретну појаву у контексту. У овом одељку испитује се други део исте хипотезе: да ли се употребом алата вештачке интелигенције, пре свега великих језичких модела, значења описана у лексикографском ресурсу могу поуздано довести у везу са синсетима Српског ворднета. Зато резултате не треба тумачити само као процену једног излаза модела, нити само као опште питање односа речника и ворднета, већ као проверу ширег методолошког питања: да ли ВЈМ, на основу речничке дефиниције, глосе синсета и проширеног контекста значења, може да предложи тип односа који је употребљив за развој и ревизију Српског ворднета?

Такво постављање задатка надовезује се на претходна истраживања о поравнању лексичко-семантичких ресурса. У тим радовима је више пута показано да поравнање значења није једноставно повезивање идентификатора, већ је у питању проблем размене информација и заједничког коришћења ресурса који различито сегментирају простор значења, различито формулишу дефиниције и различито организују семантичке односе (Matuschek 2015; Ahmadi and McCrae 2021; Ahmadi 2022). Посебно је релевантно то што Ahmadi (2022) уводи управо типове односа који су блиски схеми коришћењој у овом раду: потпуно поклапање, уже значење, шире значење, сродност и одсуство релевантне везе. Резултати добијени за српски материјал потврђују основно полазиште тих истраживања: кореспонденција између речничког значења и синсета ретко је само 1-1 и често захтева опис степена и смера семантичке блискости.

У том смислу, део резултата понаша се у складу са резултатима претходних истраживања и описаних у литератури. Као и у ранијим радовима, највећи изазови јављају се у случајевима различите грануларности, непотпуне покривености и различите функције дефиниција у два ресурса. Потпуно поклапање чини само 13,6% свих кандидатских парова, док су честе категорије сродности, ужег и ширег значења. То значи да модел не проналази само једноставне

еквиваленте, већ открива читав распон односа између речничких дефиниција и синсета. Тај налаз је у складу са претходним истраживањима, јер показује да се поравнање не може свести на питање да ли две јединице „јесу“ или „нису“ са истим значењем.

Оно што је у овом раду другачије јесте начин на који се тај проблем операционализује. За разлику од приступа који поравнање формулишу као класичан надгледани класификациони задатак над већ припремљеним златним скупом, овде се генеративни модел користи као средство за контролисано предлагање односа, уз контролу сигурности модела и кратко образложење. Другим речима, циљ није био да се добије коначан аутоматски поравнати ресурс, већ да се испита да ли велики језички модел може да смањи простор претраге, издвоји релевантне кандидате и укаже лексикографу на врсту односа који треба да провери. У односу на *H2*, то значи да се хипотеза проверава у практичном облику који је оријентисан на конкретне ресурсе: да ли је могуће конструисати поступак који на нивоу речничког значења поуздано предлаже синсете или блиске односе са синсетима, али уз задржавање људске провере као саставног дела поступка.

Модел у пуном скупу доделио позитивну ознаку за 74,0% кандидатских парова, али још важнији резултат је што су позитивне ознаке распоређене кроз више типова односа. Потпуно поклапање је само један део добијене слике; категорије сродности, ужег и ширег значења показују да описни речник и Српски ворднет често захватају близак простор значења, али га не организују на исти начин. У том смислу, резултати подржавају *H2* у ограниченом, али значајном облику: модел може да предложи релевантне везе на нивоу значења, али те везе нису увек потпуна поклапања и не могу се све непосредно претворити у стабилне односе у ресурсу.

Посебно је важно то што резултати откривају различиту грануларност два ресурса. Случајеви у којима је једна речничка дефиниција повезана са више позитивно означених синсета показују да речник понекад описује значење шире него што је оно организовано у ворднету. Обрнуто, случајеви у којима више речничких значења упућује на исти синсет показују да речник може раздвајати нијансе које у ворднету нису издвојене као засебне јединице. Ово није само грешка модела⁶³, већ структурни налаз о два ресурса. Зато је поравнање корисно и онда када не даје један коначан пар, јер открива места на којима треба преиспитати покривеност, разграничење значења и квалитет постојећих глоса.

Ручна провера првог пакета додатно прецизира овај налаз. У провереном подскупу 50,6% анотираних парова оцењено је као тачно, а још 24,1% као блиско. То значи да око три четвртине анотираних ознака које је дао модел има употребљиву вредност за редакцију, али да само половина представља ознаку која се непосредно може прихватити. Оваква расподела је у складу са општим закључком дисертације о полуаутоматском развоју ресурса: ВЈМ може знатно убрзати проналажење кандидата и усмерити пажњу лексикографа, али не може заменити експертску процену.

Коментари анотатора показују да најчешћи проблеми нису случајни, већ систематични. Прву групу проблема чине слабе или преопште семантичке везе. Оне су нарочито честе код ознаке сродности, јер модел успешно препознаје општи појмовни траг, али га понекад прецењује. Пример са лемом „адаптација” показује управо тај образац: општи појам прилагођавања није довољан да се поравнају лексикографски различита значења. Другу групу проблема чине

⁶³ Можемо имати две ситуације: 1) грешку модела: модел је предложио погрешну везу између речничког значења и синсета; 2) различиту грануларност: везе су исправне, али ресурси другачије разграничавају значења, па једном значењу у једном ресурсу одговара више значења у другом.

случајеви чине повезивања са другим значењем или другом граном значења, као код појединих парова за лему „*аберација*”, где се општи појам одступања, значења везана за астрономију и оптику, као и психичко стање не могу третирати као једноставне варијације истог односа. Трећа група проблема чине погрешно одређени типови односа између речничке дефиниције и синсета: веза међу њима постоји, али није правилно одређено о каквој се вези ради - потпуном поклапању, ужем значењу, ширем значењу или сродности.

Посебно су значајни и случајеви у којима проблем није само у моделу, већ у самом ресурсу који се упоређује. Коментари о глосама и литералима показују да поравнање може послужити као дијагностички инструмент за ревизију и допуњавање Српског ворднета и речничких извора. Тако је у Српском ворднету за лему „*агитација*“ заступљено само значење које се односи на кретање или узбуркавање, док речник бележи само значење пропагандног деловања. У корпусу су потврђена оба значења, при чему је пропагандно значење учесталије. Овај пример указује на непотпуност оба ресурса и на потребу да се њихов садржај допуни на основу корпусних потврда. Слично томе, код леме „*адресирати*“ коментар указује на проблем речничке дефиниције која спаја више аспеката радње - „писати адресу на писму“ и „упутити на нечију адресу“ који су у Српском ворднету раздвојени. Такви случајеви су методолошки важни јер показују да поравнање није само начин да се речник прикључи ворднету, већ и поступак за откривање места која треба лексикографски разјаснити.

Однос између поверења модела и поверења анотатора треба тумачити опрезно. Позитивна корелација показује да постоји извесна сагласност између тога како модел и како експерт процењују тежине случаја, али су вредности корелације сувише ниске да би поверење које даје модел могло бити употребљено као аутоматски праг за прихватање. Нарочито су важни случајеви високог поверења модела у тачност и ручне оцене нетачно. Они показују да модел може бити веома сигуран у свој погрешан закључак онда када постоји убедљива, али сувише општа или погрешно усмерена семантичка веза. Због тога поверење модела треба користити као сигнал да нечему треба дати приоритет у редакцији, а не као доказ исправности.

Образложења која даје модел, међутим, имају јасну употребну вредност, али ту вредност треба степеновати. У провереном пакету већина образложења оцењена је као корисна (65,1%), додатни део као делимично користан (28,9%), док је потпуно некорисних образложења било мало (6,0%). Њихова улога није да замене ручну проверу, већ да је учине информативнијом. Када је веза означена као тачна, образложење често брзо показује који елементи дефиниције и глосе оправдавају везу. Када је означено да веза постоји али да она није добро класификована, образложење помаже да се утврди да ли је проблем у смеру односа, степену поклапања или недовољној прецизности категорије. Чак и када је веза означена као нетачна, образложење понекад открива зашто је модел погрешно: да ли се ослонио на заједничку лему, на преопшту асоцијацију, на погрешно значење или на непотпуну глосу. Управо зато образложења треба задржати као део радног тока поравнања, али их треба третирати као помоћни дијагностички слој, не као евалуациони критеријум.

Овај аспект разликује предложени поступак од већег дела класичних приступа поравнању, у којима су одлуке најчешће засноване на особинама сличности, структурним показатељима или излазу класификатора, док је разлог појединачне одлуке остајао мање непосредно доступан. У овом раду образложење није коришћено као доказ да је модел у праву, већ као начин да се редакција људима учини бржом и садржајнијом. Тај резултат је у складу са ширим методолошким ставом дисертације: употребљивост великих језичких модела у развоју ресурса не лежи само у аутоматској одлуци, већ и у способности модела да припреми прегледан материјал с кандидатима који човек може ефикасно да провери.

Овај налаз је важан и за шири закључак о употреби великих језичких модела у развоју лексичко-семантичких ресурса, јер се надовезује на образац који се јавља у више делова дискусије. У делу о лексиконима сентимента показано је да директно пресликавање из енглеског ресурса може бити корисна полазна основа, али да не замењује адаптацију и проверу на српском материјалу. У делу о разрешењу вишезначности речи ВЈМ су најуспешнији када бирају у оквиру експлицитног инвентара значења, док грешке и ознаке за непокривена значења постају материјал за ревизију инвентара. У делу о дестилацији показано је да се део знања из полуаутоматског поступка може пренети на мање моделе, али да тај пренос остаје оперативни слој, а не замена за златну евалуацију. У преводу синсета ВЈМ су корисни за припрему кандидатских глоса и литерала, али је коначни запис синсета и даље условљен лексикографском провером. Поравнање речника и ворднета уноси исти закључак у још један део развоја ресурса: ВЈМ може да предложи тип односа и образложи га, али коначна одлука о поравнању мора остати део контролисаног, експертски провереног поступка.

Резултати такође указују на један важан правац даљег рада: формализацију финијих поткатегорија односа и грешака. Садашња петочлана шема добро раздваја основне односе, али коментари показују да би у даљој итерацији било корисно посебно означити, на пример, преопшту сродност, погрешан смер ужег односно ширег односа, различит онтолошки тип, проблем глосе у ворднету и проблем речничке дефиниције. Таква формализација не мора да буде део главног закључка изведеног из добијених резултата, али је вредна као напомена за будући рад, јер би омогућила прецизнију ревизију и јасније одлучивање о томе да ли се пар уноси као поравнање, као кандидат за ревизију или као сигнал за измену једног од ресурса.

Ограничења ових резултата морају бити јасно наглашена. Пун скуп од 3.414 парова представља излаз модела, а не ручно потврђени златни скуп. Ручно је проверен само први пакет од 100 парова, који обухвата 22 леме из почетног абецедног дела речника (16 именица, 4 глагола и 2 придева); од тога су 83 пара добила оцену анотатора. Поред тога, проверу је обавио само један анотатор, па није могуће израчунати међуанотаторску сагласност. Први пакет има висок удео позитивних ознака које је дао модел у односу на просек целог скупа, због чега га треба третирати као аналитички и дијагностички узорак, а не као репрезентативну процену целокупне тачности. За поузданију евалуацију потребно је проширити подскуп који ће бити ручно проверен, стратификовати га по ознакама модела, врстама речи, степену поверења модела и броју кандидата за одговарајући синсет, а потребно је и укључивање најмање још једног независног анотатора.

Посебно треба имати у виду да ограничење ручне провере није само у броју прегледаних парова, већ и у природи самог поступка анотације. Анотација значења захтева значајно време по јединици провере, јер анотатор мора да упореди лему, предложени синсет, дефиницију, примере употребе и положај синсета у мрежи семантичких односа. То није механичка провера исправности пара, већ лексикографска процена подударности значења. Истовремено, такву проверу не може поуздано да изврши било који учесник у поступку: потребан је анотатор са одговарајућим лексикографским искуством, познавањем структуре Српског ворднета и способношћу да разликује блиска и делимично преклопљена значења, као и контекст који их карактерише. Због тога проширење ручно провереног подскупа подразумева не само додатно време, већ и доступност довољно обучених стручних анотатора.

У односу на хипотезу *H2*, резултати поравнања подржавају тврдњу да је могуће конструисати поступак у коме велики језички модел на нивоу речничког значења предлаже синсете који су с њим подударају или су повезани неким другим односом синсети. Међутим, та подршка није исто што и доказ потпуно аутоматског поравнања. Најтачније је рећи да *H2* добија емпиријску

подршку у делу који се односи на издвајање и класификацију кандидатских веза уз подршку алата вештачке интелигенције, док коначна лексикографска интеграција захтева ручну проверу, ревизију глоса и пажљиво решавање случајева различите грануларности. У односу на хипотезу *H0*, овај резултат допуњује општи аргумент дисертације: развој Српског ворднета може се у значајној мери аутоматизовати применом великих језичких модела, али се одрживо унапређење ресурса постиже тек повезивањем обраде моделима, експлицитно представљене семантичке структуре и људске лексикографске контроле. Под експлицитном семантичком структуром подразумева се организација значења у Српском ворднету посредством синсета, глоса којима се та значења описују и непосредно забележених семантичких односа међу синсетима.

6.6 Низводни задаци и границе употребе обогаћеног ресурса

Низводни задаци у овој дисертацији немају исту методолошку улогу. Анализа сентимента служи као спољна провера слоја сентимента Српског ворднета, јер се вредности придружене синсетима непосредно користе у извођењу ознаке сентимента реченице. Циљана провера у машинском превођењу има ужу дијагностичку улогу: она испитује да ли се информација о изабраном значењу, добијена поступком разрешења вишезначности речи и повезивањем са енглеским литералима, може једноставно унети у постојећи преводилачки систем као кратак траг значења. Због те разлике резултате не треба сводити на један заједнички суд о хипотези *H4*, већ их треба тумачити као две различите провере употребљивости обогаћеног ресурса.

У анализи сентимента резултати подржавају основни смер потхипотезе *H4a*. Најбољи резултат по макро-F1 мери постиже конфигурација S7-avg, док је S7-wsd врло близу врха и има нешто вишу тачност. Обе конфигурације статистички значајно надмашују S0, односно полазни ресурс заснован на директном пресликавању из енглеског SentiWordNet-a. Овај налаз је важан јер показује да предност обогаћених вредности сентимента није видљив само у унутрашњим поређењима са синтетичким евалуационим скупом, већ и у задатку у коме се на основу тих вредности доноси одлука о сентименту целе реченице.

Истовремено, налаз треба формулисати опрезно. У односу на лексички ресурс L0, конфигурација S7-avg има бољу тачност и бољу макро-F1 меру, али разлика у макро-F1 мери после корекције за више поређења⁶⁴ остаје на граници прага статистичке значајности 0,05 (некоригована р-вредност је 0,0175, а коригована 0,0525), и не треба је тумачити као статистички значајну. Зато резултати анализе сентимента најјаче потврђују онај део хипотезе *H4a* који се односи на предност обогаћеног ресурса као мреже синсета у односу на полазно пресликавање из енглеског SentiWordNet-a, док поређење са лексичким приступом треба описати као повољно, али не и формално потврђено по примарној мери.

Однос између S7-avg и S7-wsd додатно разјашњава ограничења тренутне поставке. Конфигурација avg не бира конкретно значење у реченици, већ користи просечну вредност синсета за исту лему и врсту речи у целом лексикону сентимента, па зато има већу покривеност. Конфигурација S7-wsd, напротив, користи синсет изабран поступком разрешења вишезначности речи и зато зависи од инвентара значења изграђеног у овом раду, као и од тога колико је тај инвентар повезан са слојем сентимента Српског ворднета. Мања покривеност S7-

⁶⁴ Реч је о Холм-Бонферонијевој корекцији р-вредности за пет поређења конфигурација према макро-F1 мери. Она контролише ризик да се, због већег броја тестова, нека разлика погрешно прогласи статистички значајном. Коригују се р-вредности, а не сама макро-F1 мера.

wsd стога не указује нужно на слабост самог приступа за разрешење вишезначности речи, већ пре свега на то да инвентар и његово повезивање са ресурсом још увек нису довољно развијени.

Истовремено, разлике у тачности и макро-F1 мери између S7-avg и S7-wsd остају релативно мале, иако S7-wsd ради над ужим и захтевнијим пресеком примера. То је важан налаз за дискусију, јер показује да избор конкретног синсета већ сада даје резултат који је близак конфигурацији са већом покривеношћу. Зато је оправдано претпоставити да би даље унапређење инвентара значења и његово боље повезивање са Српским ворднетом и слојем сентимента могло повећати добитак конфигурације S7-wsd, потенцијално и у мери у којој би она надмашила S7-avg. Ипак, такав исход у овој фази треба задржати као опрезну интерпретацију важну за правац даљег рада, а не као већ потврђен закључак.

Овај резултат показује да се практична употребљивост ресурса мора процењивати кроз више мера истовремено. Макро-F1 мера је важна јер показује како се систем понаша по класама и не дозвољава да укупан резултат буде последица доминације једне класе. Тачност даје општу слику исправности предвиђања, али није толико осетљива на неравномерну расподелу класа. Покривеност, међутим, показује да ли се ресурс уопште може применити на довољно велики део текста. Зато конфигурације које имају сличне вредности макро-F1 мере и тачности могу имати различиту практичну вредност ако се значајно разликују у покривености.

Циљана провера у машинском превођењу даје другачији исход. У тој поставци није испитиван превод или формирање синсета, већ превођење реченица са српског на енглески језик уз додатну информацију о изабраном значењу циљне јединице. Резултати не показују стабилно побољшање превода након уметања напомене о значењу. У делу примера основни превод већ бира прихватљив енглески еквивалент, па додатна информација не може да донесе видљив добитак. У другим случајевима напомена о значењу доводи до парафразирања, промене стила, копирања напомене или нарушавања језичке природности реченице (енгл. *fluency*), без јасног побољшања превода циљне јединице.

Овај налаз упућује и на то да за машинско превођење нису најпогодније исте врсте евалуације као за анализу сентимента. Ако је циљ да се испита корист од разрешења вишезначности речи, потребни су циљани примери у којима је вишезначност таква да при избору најбољег решења основни преводилачки систем заиста може да погрешити. У супротном, додата информација о значењу може остати без видљивог ефекта зато што је основни превод већ довољно добар. За будуће провере зато су погоднији контрастивни скупови⁶⁵(Rios Gonzales et al. 2017), избор између више преводних еквивалената или поставке у којима се лексичко-семантичка информација уводи као ограничење, а не само као текстуална напомена.

Због тога хипотезу *H4b* у испитаној поставци треба сматрати непотврђеном. Такав исход не значи да разрешење вишезначности речи нема вредност за машинско превођење, нити да информација о значењу из ворднета не може бити корисна у преводилачким системима. Он показује уже ограничење: једноставно уметање напомене о значењу у улазни текст није довољно стабилан начин да се та информација пренесе у систем који има сопствени механизам избора преводног еквивалента. Информација о значењу у овој поставци није део архитектуре

⁶⁵ У овом контексту, контрастивни скупови су тестни скупови у којима се уз исту изворну реченицу дају исправан превод и један или више намерно измењених превода, у којима је вишезначна реч преведена у значењу које не одговара контексту. Проверава се да ли модел додељује већу вероватноћу исправном преводу него погрешним варијантама

модела, није део контролисаног декодирања и није дата као формално лексичко ограничење, већ као додатни текстуални траг. Управо зато је њен утицај посредан и непоуздан.

За тумачење хипотезе *H4* најважније је да се ова два резултата не поништавају, већ одређују домет различитих начина употребе обогаћеног Српског ворднета. Анализа сентимента показује да вредности сентимента синсета могу донети мерљиву корист када су непосредно повезане са излазном одлуком. Машинско превођење показује да иста врста информације о значењу не мора донети добитак када се спољном систему прослеђује као обична напомена у тексту. Према томе, *H4* је потврђена у ограниченом облику: подржана је у анализи сентимента, нарочито у поређењу са полазним пресликавањем из SentiWordNet-a, али није потврђена у испитаној поставци машинског превођења.

Овај исход је у складу са ширим методолошким ставом дисертације. Обогаћени лексичко-семантички ресурс није сам по себи гаранција побољшања у сваком спољном систему; његов ефекат зависи од начина на који се ресурс укључује у задатак. Када је веза између синсета и излазне одлуке непосредна, као у анализи сентимента, корист ресурса је јаснија и емпиријски видљива. Када је веза посредна, као у машинском превођењу са накнадно додатом напоменом о значењу, потребна је дубља интеграција: лексичка ограничења, контролисано декодирање, избор примера у којима основни превод заиста греш⁶⁶ или архитектура која информацију о значењу користи као саставни део процеса превођења.

Тако постављена дискусија омогућава прецизнији закључак о вредности обогаћеног Српског ворднета за низводне задатке. Резултати не подржавају јаку тврдњу да их сваки вид додавања информације из ворднета нужно побољшава. Они, међутим, подржавају умеренију и методолошки стабилнију тврдњу: обогаћени ресурс има практичну вредност када се користи у задатку чија је излазна одлука непосредно ослоњена на својства синсета, док за задатке у којима одлуку доноси независан језички систем начин интеграције ресурса постаје пресудан.

⁶⁶ Може бити корисно за евалуацију тако што проверава се да ли додата информација о значењу исправља грешке основног превода. За дораду система се анализа тих грешака може усмерити измене упита, ограничења или обуке.

7 Закључак

У овој дисертацији испитана је могућност да се методе машинског учења и велики језички модели употребе за развој, ревизију и примену Српског ворднета као лексичко-семантичког ресурса за српски језик. Полазно питање није било само да ли се постојећи ресурс може проширити новим синсетима, дефиницијама, литералима и оценама сентимента, већ и да ли се такво унапређење може искористити за већу применљивост ресурса у задацима аутоматског разумевања текста. Резултати показују да је такав приступ изводљив, али да његова поузданост зависи од контролисаног спајања излаза који даје модел, експлицитног инвентара значења, јединствене ознаке синсета, помоћу којих се исти синсет повезује са записом у инвентару, одговором модела и референтном анотацијом, ручно проверених евалуационих скупова и лексикографске ревизије.

Општа хипотеза *H0* подржана је у ограниченом, али методолошки значајном облику. Показано је да се више кључних процеса укључених у развој Српског ворднета може обавити аутоматски или полуаутоматски: додела и редакција оцена сентимента, разрешење вишезначности речи у корпусу, припрема кандидатских глоса и литерала, поравнање речничких значења са синсетима, као и формирање сребрних података за обуку мањих модела. Истовремено, резултати не подржавају јаку тврдњу о потпуно самосталној аутоматизацији развоја ресурса. Најстабилнији модел рада остаје полуаутоматски: велики језички модели убрзавају, проширују и усмеравају лексикографски рад, али не замењују експертску проверу и методолошку контролу.

Хипотеза *H1*, која се односи на слој сентимента Српског ворднета, добила је емпиријску подршку на више нивоа евалуације. Резултати показују да непосредно пресликавање оцена из енглеског SentiWordNet-а представља корисну полазну поставку, али није довољно као завршни слој сентимента за српски језик. Лексикони добијени поступцима развијеним у овој дисертацији боље разликују објективне, слабо поларизоване и јаче поларизоване синсете од полазног пресликавања. Посебно се издвајају лексикони *S5* и *S7* као погодни кандидати за степеновано моделовање сентимента, док је *S7* добио најјачу потврду у пракси при екстерној евалуацији анализе сентимента реченица. Због тога се *S7* може посматрати као најрелевантнији кандидат за практично употребљив ресурс за придруживање оцена сентимента синсетима у Српском ворднету. Ови резултати показују да унапређење Српског ворднета није само у проширењу ресурса, већ и у повећању његове применљивости у конкретном низводном задатку.

У делу рада посвећеном разрешењу вишезначности речи показано је да Српски ворднет може постати оперативна основа за аутоматско обележавање текстуалних појава синсетима. Формиран је почетни инвентар значења WSD-sr-sens-inv-v1, а затим и његова ревидирана верзија WSD-sr-sens-inv-v2, настала на основу аутоматске анотације, ручне провере и анализе случајева у којима постојећи инвентар није обезбеђивао одговарајуће значење. У том процесу разрешење вишезначности речи није било само примена постојећег ресурса, већ и механизам његове ревизије: грешке, недодељени случајеви и ознаке за непокривена значења указивали су на места на којима је потребно проширити инвентар значења или у њему прецизније разграничити постојеће записе.

Посебан допринос представља формирање ручно провереног златног скупа ELEXIS-WSD-sr-gold, са развојно-аналитичким делом и издвојеним делом за тестирање. Оваква организација омогућила је да се један део материјала користи за анализу грешака и ревизију инвентара значења, док други део служи као стабилнија основа за процену успешности поступка.

Резултати показују да велики језички модели могу успешно да се користе за контролисан избор значења када је избор ограничен на унапред дефинисан скуп кандидата из инвентара синсета. Тиме је хипотеза *H2* подржана у делу који се односи на поуздано придруживање текстуалних појава синсетима Српског ворднета.

Други део *H2* испитан је кроз лексичко-семантичко поравнање описног речника и Српског ворднета. Резултати показују да велики језички модел може да предложи корисне кандидатске везе између речничких значења и синсета, али и да такве везе нису увек потпуна поклапања. Честе су категорије сродности, ужег и ширег значења, што потврђује да речник и ворднет не сегментирају простор значења на исти начин. Ручна провера првог пакета показала је да излаз модела има значајну дијагностичку вредност за ревизију, али да није довољан за потпуно аутоматско поравнање. Због тога *H2* треба сматрати подржаном у делу који се односи на издвајање и класификацију кандидатских веза уз подршку алата вештачке интелигенције, али не као доказ успешности коначног аутоматског поравнања ресурса.

Хипотеза *H3*, која се односи на генеративну допуну и усклађивање синсета, такође је делимично подржана. У експерименту са преводом синсета показано је да велики језички модели могу да припреме глосе које су више у духу српског језика као и употребљивији почетни скуп литерала од непосредног превођења. Најбоље резултате дао је концептуални радни ток, у коме модел најпре стабилизује значење синсета, а тек затим приступа формулисању српске глосе и избору литерала. Овај налаз је важан јер показује да квалитет записа синсета не зависи само од језичке природности превода већ и од тога да ли је значење претходно довољно јасно концептуализовано и разграничено.

Истовремено, резултати превода синсета не подржавају потпуно аутоматско проширивање Српског ворднета без експертске провере. У сва три испитана радна тока јављају се литерали који недостају или су сувишни, као и случајеви у којима предложене лексеме имају слично али не и исто значење, или оне које се не могу прихватити јер не припадају српском језику (нпр. јер их је модел формирао прилагођавањем из енглеског). Зато је најпрецизнији закључак да су велики језички модели корисни за припрему преводних кандидата, али да коначна одлука о глоси, литералима и односу према постојећој структури Српског ворднета мора остати део контролисаног лексикографског поступка.

Резултати дестилације показују да се део знања добијеног полуаутоматском анотацијом може пренети на мање и оперативно једноставније моделе. Сребрни скуп нема статус златног стандарда, али има употребну вредност као материјал за обучавање модела који могу да рангирају кандидатска значења и служе као помоћни слој у обради већих количина текста. Најуспешнији дестиловани модели не достижу резултате најбољих генеративних поставки, али њихов значај није у замени најтачнијег поступка, већ у могућности да се успостави ефикаснији радни ток: брза припремна обрада, издвајање тежих случајева, упоређивање одлука модела и усмеравање провере коју људи треба да обаве.

Хипотеза *H4* потврђена је само у ограниченом облику. Резултати анализе сентимента подржавају *H4a*, пре свега кроз предност конфигурација *S7-avg* и *S7-wsd* у односу на полазни ресурс *S0*. У ресурсу *S0* вредности сентимента непосредно су пренете из енглеског SentiWordNet-а на одговарајуће синсете Српског ворднета. Обогаћене вредности сентимента нису донеле само интерно побољшање ресурса, већ су довеле до мерљиве користи у задатку у коме се на основу вредности синсета доноси одлука о сентименту реченице. У поређењу са лексичким приступом резултате треба формулисати опрезније, јер предност није подједнако формално потврђена по свим мерама. Ипак, овај део евалуације показује да Српски ворднет,

када се обогати и прилагоди српском језичком материјалу, може бити применљивији у анализи текста.

Насупрот томе, *H4b* није потврђена у испитаној поставци машинског превођења. Једноставно уметање напомене о значењу, засноване на изабраном значењу и енглеским литералима Принстонског ворднета, није довело до стабилног побољшања превода. Овај налаз не поништава вредност слоја за разрешење вишезначности речи, већ прецизира границе његове примене. Када је веза између синсета и излазне одлуке непосредна, као у анализи сентимента, корист ресурса је емпиријски видљивија. Када се информација о значењу уводи у преводилачки систем посредно, као обична текстуална напомена, њен ефекат зависи од начина интеграције. За машинско превођење су зато потребне поставке које омогућавају експлицитна лексичка ограничења, контролисано декодирање или другачији архитектонски начин употребе информације о значењу.

У целини, резултати ове дисертације показују да се применом метода машинског учења и великих језичких модела, уз ручну проверу и лексикографску редакцију, може унапредити Српски ворднет и повећати његова употребљивост у задацима аутоматског разумевања текста. Најјача потврда применљивости добијена је у анализи сентимента и у аутоматском обележавању текстуалних појава синсетима. Додатни резултати, као што су превод синсета, лексичко-семантичко поравнање и дестилација, показују како се појединачни делови поступка могу повезати у шири радни ток развоја ресурса. Главни допринос рада је зато доказ да велики језички модели не могу самостално да изграде или ревидирају ворднет, али могу знатно да убрзају и усмере развој српског лексичко-семантичког ресурса када су укључени у проверљив, ограничен и лексикографски надгледан поступак.

7.1 Доприноси

Научни доприноси дисертације могу се сажети у неколико целина. Први допринос односи се на изградњу и евалуацију слоја сентимента Српског ворднета, укључујући лексиконе сентимента настале у овом раду и моделе који из глоса синсета израчунавају оцене сентимента. Овај допринос треба разликовати од спољних полазних ресурса и поредбених поставки, као што су *Senti-pol-sr* и *S0* заснован на пресликавању из *SentiWordNet-a*. У раду је показано да непосредно преузимање оцена из *SentiWordNet-a* није довољно за српски језик и да поступци развијени у дисертацији могу да произведу стабилније оцене сентимента на нивоу синсета. Нарочито је значајан лексикон *S7*, јер је у екстерној евалуацији показао најјачу практичну потврду и може служити као основа за унос, односно придруживање вредности сентимента синсетима у Српском ворднету. Допринос тиме није ограничен на придруживање коначних вредности лексикону, већ обухвата и поступак модела за њихово добијање.

Други допринос односи се на развој поступка за разрешење вишезначности речи у српском корпусном материјалу. Формирањем инвентара значења *WSD-sr-sens-inv-v1* и *WSD-sr-sens-inv-v2* успостављен је радни слој који повезује корпусне појаве, кандидате за значења, дефиниције, идентификаторе синсета и Српски ворднет. Овим је омогућено да се појаве лексема у тексту систематски повезују са синсетима, али и да се грешке у анотацији користе као сигнал за даљу ревизију ресурса.

Трећи допринос представља ручно проверени златни скуп *ELEXIS-WSD-sr-gold*, који омогућава евалуацију поступка за разрешење вишезначности речи на српском језику. Посебна вредност овог скупа је у раздвајању развојно-аналитичког дела и издвојеног дела за тестирање, чиме се јасније разликују подаци који служе за ревизију ресурса од података који служе за процену успешности поступка.

Четврти допринос односи се на испитивање сребрног скупа и дестилације мањих модела. Резултати показују да полуаутоматски добијени подаци могу имати практичну вредност за обуку ефикаснијих модела, али и да се њихова употреба мора јасно разграничити од златне евалуације. На тај начин је предложен модел рада у коме се скупа и тачнија генеративна анотација, сребрни подаци, дестиловани модели и ручна провера повезују у један итеративан поступак.

Пети допринос односи се на испитивање превода синсета помоћу великих језичких модела. Показано је да радни ток који најпре стабилизује концептуални опис синсета даје боље резултате од непосредног превођења глосе и литерала. Ово је важан методолошки налаз, јер показује да проширивање ворднета не треба третирати као обичан преводни задатак, већ као лексичко-семантичко формирање записа кандидата за синсете.

Шести допринос односи се на лексичко-семантичко поравнање описног речника и Српског ворднета. У раду је показано да велики језички модели могу да предложе тип односа између речничког значења и синсета, као и да образложење модела може имати дијагностичку вредност за редакцију експерата. Истовремено је показано да различита грануларност речника и ворднета захтева финију анализу односа од једноставног прихватања или одбацивања поравнања.

Седми допринос представљају скупови кандидата који су настали као резултат ових поступака. То се односи на записе кандидата за синсете добијене при преводу и допуни синсета, кандидате веза добијене у поравнању описног речника и Српског ворднета, као и на поступке и моделе којима се такви кандидати припремају, рангирају или проверавају. Њихова вредност није у томе што се могу непосредно унети у Српски ворднет без провере, већ у томе што лексикографски рад усмеравају ка јединицама које имају највећи потенцијал за проширивање, редиговање или повезивање ресурса.

Коначно, допринос дисертације је и у томе што повезује више наизглед одвојених поступака у јединствену методолошку слику. Слој сентимента, разрешење вишезначности речи, превод синсета, поравнање ресурса, сребрни подаци и дестилација нису независни експерименти, већ делови истог ширег питања: како учинити да Српски ворднет буде богатији, поузданији и применљивији у задацима обраде природног језика.

7.2 Ограничења и даљи рад

Резултате дисертације треба тумачити у светлу више ограничења. Изграђени ручно проверени скуповису методолошки важни, али је њихов обим и даљеограничен. Ово је нарочито важно за превод синсета и лексичко-семантичко поравнање, где су детаљно проверени мањи подскупови и где је у појединим фазама учествовао само један анотатор. Због тога будући рад треба да прошири ручно проверене податке, укључи додатне анотаторе и омогући проверу сагласности између анотатора.

У области слоја сентимента потребно је проширити и стратификовати евалуацију. Посебно је важно проверити да ли се налази добијени на синтетичком евалуационом скупу и скупу од 3000 новинских реченица преносе на друге домене српског језика. Даљи рад треба да обухвати калибрацију прагова за позитивну и негативну класу, анализу по врстама речи и боље повезивање слоја сентимента са слојем за разрешење вишезначности речи. Посебан правац представља повећање покривености конфигурације која користи изабрани синсет, јер тренутни резултати показују да је ограничење више у покривености инвентара и његовој повезаности са слојем сентимента него у самој идеји моделовања сентимента синсета.

У области разрешења вишезначности речи потребно је наставити итеративну обраду српског дела корпуса ELEXIS-WSD и постепено проширивати златни скуп. Посебну пажњу треба посветити лексемама са великим бројем кандидатских значења, затим глаголима, вишечланим изразима и уопште случајевима у којима су дефиниције блиске или недовољно разграничене. Уместо ослањања искључиво на веће опште скупове, корисно је формирати мање циљане евалуационе скупове за најтеже типове примера, јер управо они најбоље показују где инвентар значења и Српски ворднет треба даље ревидирати.

Један од важних праваца даљег рада јесте ближе повезивање инвентара значења са самим Српским ворднетом. У овој дисертацији инвентар значења имао је улогу радног слоја који омогућава контролисано разрешење вишезначности речи. У наредним фазама треба испитати у којој мери се тај слој може директније интегрисати са структуром ворднета тако да нова значења, исправљене дефиниције, примери употребе и сигнали добијени преко ознака за непокривена значења постану део контролисаног поступка ревизије Српског ворднета, а не само елемент одвојеног радног инвентара.

На тај правац се надовезује употреба добијених скупова кандидата за стварно проширење Српског ворднета. Кандидати за саме записе синсета, кандидати за успостављање везе из поравнања са описним речником и случајеви означени као непокривена значења могу послужити као полазни материјал за додавање нових синсета, прецизирање постојећих глоса и уређивање односа између значења. Након лексикографске провере и прихватања таквих синсета, потребно је испитати и поступак доделе оцена сентимента новим синсетима, ослањајући се на најуспешнији лексикон и поступак заснован на језичким моделима из ове дисертације, пре свега на лексикон S7. Тако би се проширење основног ресурса одмах повезало са проширењем слоја сентимента. У том смислу, будући рад треба да развије моделе и радне токове који не само да предлажу кандидате за проширење, већ их повезују са постојећом структуром Српског ворднета и са одговарајућим вредностима сентимента.

Посебно значајан правац даљег рада јесте повезивање резултата превода синсета са поступком разрешења вишезначности речи. Концептуални радни ток, који је у преводу синсета дао најбоље резултате, могао би се употребити за обогаћивање приказа значења кандидата у WSD поступку. Уместо да модел бира значење само на основу постојеће кратке глосе из ворднета, треба проверити да ли проширени концептуални опис синсета, добијен контролисаним радним током, омогућава боље разграничење блиских значења у контексту. Овај правац је важан јер повезује два налаза дисертације: да је квалитетна концептуализација синсета корисна за формирање глоса и да је јасно представљање значења кандидата пресудно за успешно разрешење вишезначности речи.

Још један правац даљег рада јесте примена поступка разрешења вишезначности речи на саме глосе Српског ворднета. Уместо да се глоса посматра само као кратак текстуални опис синсета, могуће је аутоматски означити лексеме које се у њој јављају синсетима којима припадају. На тај начин би се добио граф у коме се један синсет повезује са другим синсетима који су активирани у његовој глоси. Имајући у виду да је ово врло захтеван задатак, спроводио би се итеративно, почевши од првог иницијалног скупа концепата са синсетима из BCS=1 (~1300 синсета), потом други и трећи са по око 3500 синсета. Овај правац се надовезује на поступке описане у прегледу истраживања, где је у изградњи SentiWordNet-а искоришћена чињеница да су речи у глосама Принстонског ворднета означене идентификаторима синсета, а затим је над тако добијеним графом примењена анализа заснована на насумичној шетњи. За Српски ворднет би сличан поступак могао да има двоструку улогу: да омогући анализу графа повезаности синсета и да открије глосе у којима су појмови из дефиниције нејасни, недовољно

разграничени или потенцијално погрешно повезани. Такав граф би се могао користити и као додатни сигнал за редакцију синсета, за проверу семантичких односа и, у каснијој фази, за стабилизацију или ширење оцена сентимента.

У области лексичко-семантичког поравнања потребно је развити финије категорије односа и грешака.

Даљи рад треба да обухвати и развој агентског радног тока за редакцију, експанзију и унапређење ресурса. Уместо једног великог промпта или једног излаза модела, поступак би могао бити организован као низ мањих проверљивих корака: претрага постојећих синсета, анализа глосе, провера литерала, поређење са Принстонским ворднетом, провера речничких дефиниција, издвајање корпусних примера и класификација спорних случајева. У таквом окружењу агент не би доносио коначну лексикографску одлуку, већ би позивао специјализоване алате, бележио разлоге за прихватање или одбацавање кандидата и припремао материјал за ревизију коју треба да обаве људи.

У области дестилације и сребрних података будући рад треба да испита активније облике избора примера за ручну проверу. Посебно су вредни случајеви у којима се дестиловани модел, генеративни модел и ручно правило не слажу, као и случајеви у којима је разлика између најбоље ранжираних значења кандидата мала. Такви примери могу бити основа за активније машинско учење и економичнију употребу људског рада, јер највише доприносе разјашњавању граница између блиских значења.

Коначно, у низводним задацима потребно је испитати дубље начине интеграције информације о значењу. Резултати машинског превођења показују да једноставно додавање напомене о значењу није довољно стабилно. Будући експерименти треба да користе контрастивне скупове, вишезначне примере код којих основни систем заиста може да погреши због вишезначности, као и системе који подржавају експлицитна лексичка ограничења или контролисано декодирање. Тако би се прецизније испитало у којим условима слој за разрешење вишезначности речи може да допринесе превођењу.

Општи правац даљег рада, према томе, није само повећање количине података, већ прелазак са појединачних експеримената на интегрисан, агентски подржан и лексикографски контролисан систем за развој Српског ворднета. Такав систем би објединио разрешење вишезначности речи, превод и допуну синсета, лексичко-семантичко поравнање, анализу сентимента, дестилацију мањих модела и ручну ревизију. На тај начин би се наставио основни методолошки смер ове дисертације: раст и обогаћивање Српског ворднета као ресурса нису довољни, већ треба обезбедити и да његов развој буде одржив, а његов састав проверљив и применљив у задацима аутоматског разумевања текста на српском језику.

8 Литература

- Agirre, Eneko, Oier López de Lacalle, and Aitor Soroa. 2014. “Random Walks for Knowledge-Based Word Sense Disambiguation.” *Computational Linguistics* 40 (1): 57–84. https://doi.org/10.1162/COLI_a_00164.
- Ahmadi, Sina. 2022. “Monolingual Alignment of Word Senses and Definitions in Lexicographical Resources.” <https://arxiv.org/abs/2209.02465>.
- Ahmadi, Sina, and John P. McCrae. 2021. “Monolingual Word Sense Alignment as a Classification Problem.” In *Proceedings of the 11th Global Wordnet Conference*, edited by Piek Vossen and Christiane Fellbaum. Global Wordnet Association. <https://doi.org/10.18653/v1/2021.gwc-1.9>.
- Al Tarouti, Feras, and Jugal Kalita. 2016. “Enhancing Automatic Wordnet Construction Using Word Embeddings.” In *Proceedings of the Workshop on Multilingual and Cross-Lingual Methods in NLP*, edited by Dipanjan Das, Chris Dyer, Manaal Faruqui, and Yulia Tsvetkov. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-1204>.
- Al-Edhari, Ammar, and Mohsen Kahani. 2025. “LughaNet: Automated Arabic WordNet Construction and Evaluation Using Semantic Question Similarity.” *Journal of Soft Computing and Data Mining* 6 (1): 107–26.
- Alhazmi, Samah, and William Black. 2013. “Arabic SentiWordNet in Relation to SentiWordNet 3.0.” *John McNaught International Journal of Computational Linguistics (IJCL)*, no. 4: 1.
- Almeman, Fatemah Yousef, Steven Schockaert, and Luis Espinosa Anke. 2024. “WordNet under Scrutiny: Dictionary Examples in the Era of Large Language Models.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.1538/>.
- Association for Computational Linguistics. 2026. “ACL Anthology.” <https://aclanthology.org>.
- Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, et al. European Language Resources Association (ELRA). <https://aclanthology.org/L10-1531/>.
- Basile, Pierpaolo, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. “Exploring the Word Sense Disambiguation Capabilities of Large Language Models.” *arXiv Preprint arXiv:2503.08662*, ahead of print. <https://doi.org/10.48550/arXiv.2503.08662>.
- Bergh, Johann, Jörg Waitelonis, and Melanie Siegel. 2025. “Leveraging LLMs for Constructing WordNets Automatically as Bilingual Resources.” In *Proceedings of the 13th Global Wordnet Conference*, edited by Chiara Zanchi, Luca Brigada Villa, Erica Biagetti, Alexandre Rademaker, Francis Bond, and German Rigau. Global Wordnet Association. <https://doi.org/10.18653/v1/2025.gwc-1.24>.

- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. 2021. “On the Opportunities and Risks of Foundation Models.” Version 3. Preprint, arXiv. <https://doi.org/10.48550/ARXIV.2108.07258>.
- Bond, Francis, and Ryan Foster. 2013. “Linking and Extending an Open Multilingual Wordnet.” *Proceedings of the 51st Annual Meeting of the ACL (Volume 1: Long Papers)*, 1352–62. <https://aclanthology.org/P13-1133/>.
- Bond, Francis, Michael Wayne Goodman, Ewa Rudnicka, Luis Morgado da Costa, Alexandre Rademaker, and John P. McCrae. 2023. “Documenting the Open Multilingual Wordnet.” In *Proceedings of the 12th Global Wordnet Conference*, edited by German Rigau, Francis Bond, and Alexandre Rademaker. Global Wordnet Association. <https://doi.org/10.18653/v1/2023.gwc-1.18>.
- Bond, Francis, and Kyonghee Paik. 2012. “A Survey of Wordnets and Their Licenses.” *Proceedings of the 6th Global WordNet Conference (GWC 2012)* (Matsue, Japan), 64–71.
- Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. “Language Models Are Few-Shot Learners.” In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, vol. 33. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- Camacho-Collados, Jose, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. 2019. “SenseDefs: A Multilingual Corpus of Semantically Annotated Textual Definitions: Exploiting Multiple Languages and Resources Jointly for High-Quality Word Sense Disambiguation and Entity Linking.” *Language Resources and Evaluation* 53 (2): 251–78. <https://doi.org/10.1007/s10579-018-9421-3>.
- Camacho-Collados, Jose, and Mohammad Taher Pilehvar. 2018. “From Word to Sense Embeddings: A Survey on Vector Representations of Meaning.” *Journal of Artificial Intelligence Research* 63: 743–88.
- Cao, Zhihan, Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2025. “Misalignment of Semantic Relation Knowledge between WordNet and Human Intuition.” *Proceedings of the 13th Global Wordnet Conference*, 25–36.
- Carpuat, Marine, and Dekai Wu. 2005. “Word Sense Disambiguation vs. Statistical Machine Translation.” In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, edited by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer. Association for Computational Linguistics. <https://doi.org/10.3115/1219840.1219888>.
- Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. “Word Sense Disambiguation Improves Statistical Machine Translation.” In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, edited by Annie Zaenen and Antal van den Bosch. Association for Computational Linguistics. <https://aclanthology.org/P07-1005/>.
- Čibej, Jaka, Ranka Stanković, Ana Ostroški Anić, Simon Krek, and Carole Tiberius. 2025. “The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene.” *Proceedings of the South Slavic Languages in Digital Environment—JuDig*, 45–59.

- Comanici, Gheorghe, Eric Bieber, Mike Schaekermann, et al. 2025. “Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.” arXiv:2507.06261. Preprint, arXiv, December 19. <https://doi.org/10.48550/arXiv.2507.06261>.
- Cruse, David A. 2001. *Lexical Semantics*. Reprint. Cambridge Textbooks in Linguistics. Cambridge Univ. Press.
- Das, Amitava, and Sivaji Bandyopadhyay. 2010. “SentiWordNet for Indian Languages.” In *Proceedings of the Eighth Workshop on Asian Language Resources*, edited by Sarmad Hussain, Virach Sornlertlamvanich, and Hammam Riza. Coling 2010 Organizing Committee. <https://aclanthology.org/W10-3208/>.
- Dehkharghani, Rahim, Y. Saygin, B. Yanikoglu, and Kemal Oflazer. 2016. “SentiTurkNet: A Turkish Polarity Lexicon for Sentiment Analysis.” *Language Resources and Evaluation* 50 (3): 667–85. <https://doi.org/10.1007/s10579-015-9307-6>.
- Denecke, Kerstin. 2008. “Using SentiWordNet for Multilingual Sentiment Analysis.” *International Conference on Data Engineering*, 507–12. <https://doi.org/10.1109/ICDEW.2008.4498370>.
- Denecke, Kerstin, and Daniel Reichenpfader. 2023. “Sentiment Analysis of Clinical Narratives: A Scoping Review.” *Journal of Biomedical Informatics* 140: 104336. <https://doi.org/10.1016/j.jbi.2023.104336>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.
- Esuli, Andrea, and Fabrizio Sebastiani. 2006. “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.” In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, edited by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, et al. European Language Resources Association (ELRA). <https://aclanthology.org/L06-1225/>.
- Fan, Angela, Shruti Bhosale, Holger Schwenk, et al. 2021. “Beyond English-Centric Multilingual Machine Translation.” *Journal of Machine Learning Research* 22 (107): 1–48.
- Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. <https://books.google.com/books?id=Rehu8OOzMIMC>.
- Fellbaum, Christiane. 2005. “WordNet and Wordnets.” In *Encyclopedia of Language and Linguistics, Second Edition*, edited by Keith Brown. Elsevier.
- Filip, David, Miloš Jakubiček, Simon Krek, John McCrae, and Michal Měchura, eds. 2025. *Data Model for Lexicography (DMLex) Version 1.0*. OASIS Open. <https://docs.oasis-open.org/lexidma/dmlex/v1.0/os/dmlex-v1.0-os.html>.
- Formentelli, Maicol, Sabrina Cerini, Valentina Compagnoni, Alice Demontis, and Caterina Gandini. 2007. “Micro-WNOp: A Gold Standard for the Evaluation of Automatically Compiled Lexical

Resources for Opinion Mining.” In *Language Resources and Linguistic Theory*, edited by Andrea Sansò. Franco Angeli.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 2025. “Gemini: A Family of Highly Capable Multimodal Models.” arXiv:2312.11805. Preprint, arXiv, May 9. <https://doi.org/10.48550/arXiv.2312.11805>.

Global WordNet Association. n.d. “Wordnets in the World.” Accessed March 18, 2026. <https://globalwordnet.github.io/resources/wordnets-in-the-world>.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. “The Llama 3 Herd of Models.” arXiv:2407.21783. Preprint, arXiv, November 23. <https://doi.org/10.48550/arXiv.2407.21783>.

Gromann, Dagmar, Elena-Simona Apostol, Christian Chiarcos, et al. 2024. “Multilinguality and LLOD: A Survey across Linguistic Description Levels.” *Semantic Web* 15 (5): 1915–58. <https://doi.org/10.3233/SW-243591>.

Guo, Ping, Yubing Ren, Yue Hu, et al. 2024. “Teaching Large Language Models to Translate on Low-Resource Languages with Textbook Prompting.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.1362/>.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. “Distilling the Knowledge in a Neural Network.” arXiv:1503.02531. Preprint, arXiv, March 9. <https://doi.org/10.48550/arXiv.1503.02531>.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80. <https://doi.org/10.1162/neco.1997.9.8.1735>.

Indig, Balázs, and Márton Miháltz. n.d. “Python3 API for WordNet XML (Hungarian WordNet in BalkaNet/VisDic XML Format).” <https://github.com/ppke-nlpg/pywnxml>.

Ji, Ziwei, Nayeon Lee, Rita Frieske, et al. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55 (12): 1–38. <https://doi.org/10.1145/3571730>.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, et al. 2023. “Mistral 7B.” arXiv:2310.06825. Preprint, arXiv, October 10. <https://doi.org/10.48550/arXiv.2310.06825>.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, et al. 2018. “Marian: Fast Neural Machine Translation in C++.” In *Proceedings of ACL 2018, System Demonstrations*, edited by Fei Liu and Thamar Solorio. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-4020>.

Jurafsky, Daniel, and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd ed. <https://web.stanford.edu/~jurafsky/slp3/>.

Kahusk, Neeme. n.d. “Eurown: An EuroWordNet Module for Python.” https://www.cfilt.iitb.ac.in/gwc2010/pdfs/63_eurown_for_python_Kahusk.pdf.

- Khan, Fahad, Christian Chiarcos, Thierry Declerck, et al. 2022. "A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data." In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, edited by Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, and Maxim Ionov. European Language Resources Association. <https://aclanthology.org/2022.ldl-1.9/>.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, et al. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1 (1): 7–36. <https://doi.org/10.1007/s40607-014-0009-9>.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. "The INCEPTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (Santa Fe, New Mexico), August, 5–9. <https://aclanthology.org/C18-2002/>.
- Koeva, Svetla, Cvetana Krstev, and Duško Vitas. 2008. "Morpho-Semantic Relations in WordNet: A Case Study for Two Slavic Languages." *Proceedings of the Fourth Global WordNet Conference* (Szeged), 239–54.
- Krstev, Cvetana. 2008. *Processing of Serbian: Automata, Texts, and Electronic Dictionaries*. Faculty of Philology, University of Belgrade.
- Krstev, Cvetana, Svetla Koeva, and Duško Vitas. 2006. "Towards the Global Wordnet." *Conference Abstracts of the First International Conference of Digital Humanities Organisations (ADHO) Digital Humanities 2006* (Paris-Sorbonne), July 5, 114–17. <http://poincare.matf.bg.ac.rs/~cvetana/biblio/DigHum06-KKV.pdf>.
- Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. "Using Textual and Lexical Resources in Developing Serbian Wordnet." *Romanian Journal of Information Science and Technology* 7 (1–2): 147–61.
- Krstev, Cvetana, and Ranka Stanković. 2022. "Report on the Serbian Language D1.35." In *European Language Equality: WPI: European Language Equality – Status Quo in 2020/2021 (Task 1.3 Language Technology Support of Europe's Languages in 2020/2021)*, edited by Maria Giagkou, Stelios Piperidis, Georg Rehm, and Jane Dunne. ELE Consortium. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_35___Language_Report_Serbian_.pdf.
- Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality*, edited by Georg Rehm and Andy Way. Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7_32.
- Krstev, Cvetana, Ranka Stanković, and Aleksandra Marković. 2024. "Annotation of MWEs and NEs in the Serbian Extension of ELEXIS-WSD: Comparisons, Solutions and Open Questions." *2nd UniDive Workshop in Naples*, 1–4.
- Krstev, Cvetana, Ranka Stanković, Aleksandra Marković, and Milica Ikonić Nešić. 2025. "Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository." *3rd General Meeting Hungarian Research Centre for Linguistics, Budapest, Hungary 29-30 January 2025*, 4.

- Krstev, Cvetana, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. "Towards the Semantic Annotation of SR-ELEXIS Corpus: Insights into Multiword Expressions and Named Entities." In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, edited by Archana Bhatia, Gosse Bouma, A. Seza Doğruöz, et al. ELRA and ICCL. <https://aclanthology.org/2024.mwe-1.15/>.
- Kwong, Oi Yee. 1998. "Aligning WordNet with Additional Lexical Resources." *Usage of WordNet in Natural Language Processing Systems*. <https://aclanthology.org/W98-0710/>.
- Lam, Khang Nhut, Feras Al Tarouti, and Jugal Kalita. 2014. "Automatically Constructing Wordnet Synsets." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Kristina Toutanova and Hua Wu. Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-2018>.
- Lesk, Michael E. 1986. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86)* (Toronto, Canada), 24–26. <https://doi.org/10.1145/318723.318728>.
- Liu, Bing. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing, Second Edition*, edited by Fred J. Damerau and Nitin Indurkha. Chapman & Hall/CRC.
- Liu, Frederick, Han Lu, and Graham Neubig. 2018. "Handling Homographs in Neural Machine Translation." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, edited by Marilyn Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1121>.
- Liu, Nelson F., Kevin Lin, John Hewitt, et al. 2024. "Lost in the Middle: How Language Models Use Long Contexts." *Transactions of the Association for Computational Linguistics* 12 (February): 157–73. https://doi.org/10.1162/tacl_a_00638.
- Ljubešić, Nikola, and Davor Lauc. 2021. "BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian." In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, edited by Bogdan Babych, Olga Kanishcheva, Preslav Nakov, et al. Association for Computational Linguistics. <https://aclanthology.org/2021.bsnlp-1.5/>.
- Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar, and José Camacho-Collados. 2021. "Analysis and Evaluation of Language Models for Word Sense Disambiguation." *Computational Linguistics* 47 (2): 387–443. https://doi.org/10.1162/coli_a_00405.
- Madaan, Aman, Niket Tandon, Prakhar Gupta, et al. 2023. "Self-Refine: Iterative Refinement with Self-Feedback." *Advances in Neural Information Processing Systems* 36: 46534–94.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge university press.
- Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* (Cambridge, MA) 47 (2): 255–308. https://doi.org/10.1162/coli_a_00402.

- Martelli, Federico, Roberto Navigli, Simon Krek, et al. 2021. “Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages.” *Proceedings of the eLex 2021 Conference*, 377–95.
- Martelli, Federico, Roberto Navigli, Simon Krek, et al. 2023. “Parallel Sense-Annotated Corpus ELEXIS-WSD 1.1.” Slovenian Language Resource Repository CLARIN.SI. <https://www.clarin.si/repository/xmlui/handle/11356/1799>.
- Masethe, Hlaudi Daniel, Mosima Anna Masethe, Sunday Olusegun Ojo, Fausto Giunchiglia, and Pius Adewale Owolawi. 2024. “Word Sense Disambiguation for Morphologically Rich Low-Resourced Languages: A Systematic Literature Review and Meta-Analysis.” *Information* 15 (9). <https://doi.org/10.3390/info15090540>.
- Matuschek, Michael. 2015. *Word Sense Alignment of Lexical Resources*. <https://doi.org/10.26083/TUPRINTS-00004355>.
- Meconi, Domenico, Simone Stirpe, Federico Martelli, Leonardo Lavallo, and Roberto Navigli. 2025. “Do Large Language Models Understand Word Senses?” In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, edited by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1720>.
- Medagoda, Nishantha, Subana Shanmuganathan, and Jacqueline Whalley. 2015. “Sentiment Lexicon Construction Using SentiWordNet 3.0.” *2015 11th International Conference on Natural Computation (ICNC)*, 802–7. <https://doi.org/10.1109/ICNC.2015.7378094>.
- Meta. 2025. “The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation.” <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” Version 3. Preprint, arXiv. <https://doi.org/10.48550/ARXIV.1301.3781>.
- Miller, George A. 1995. “WordNet: A Lexical Database for English.” *Communications of the ACM* 38 (11): 39–41. <https://doi.org/10.1145/219717.219748>.
- Mišić, Ksenija, Sara Anđelić, Lenka Ilić, Dajana Osmani, Milica Manojlović, and Dušica Filipović Đurđević. 2024. “AN OPEN DATABASE OF SENSES FOR SERBIAN POLYSEMOUS NOUNS, VERBS, AND ADJECTIVES.” *STUDIES IN LANGUAGE AND MIND*, no. 5 (May): 7–28. <https://doi.org/10.19090/slm.5.1>.
- Mistral AI. 2025. “Mistral Small 3.2.” <https://docs.mistral.ai/models/mistral-small-3-2-25-06>.
- Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2015. “Hybrid Sentiment Analysis Framework for a Morphologically Rich Language.” *Journal of Intelligent Information Systems* 46 (3): 599–620. <https://doi.org/10.1007/s10844-015-0372-5>.
- Mohanty, Gaurav, A. Kannan, and R. Mamidi. 2017. *Building a SentiWordNet for Odia*. 143–48. <https://doi.org/10.18653/v1/W17-5219>.
- Navigli, Roberto. 2009. “Word Sense Disambiguation: A Survey.” *ACM Comput. Surv.* (New York, NY, USA) 41 (2). <https://doi.org/10.1145/1459352.1459355>.

- Obare, Stephen, Abejide Ade-Ibijola, and Kennedy Ogada. 2025. "Developing A Dynamic WordNet for Under-Resourced Languages." *Journal of Information Science & Engineering* 41 (5).
- Obradović, Ivan, Cvetana Krstev, Gordana Pavlović-Lazetić, and Duško Vitas. 2004. "Corpora Based Validation of WordNet Using Frequency Parameters." *Proceedings of the Second Global WordNet Conference (GWC 2004)* (Brno, Czech Republic), 181–86.
- Oliver, Antoni. 2020. "Aligning Wikipedia with WordNet: A Review and Evaluation of Different Techniques." *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (Marseille, France), 4851–58. <https://aclanthology.org/2020.lrec-1.597>.
- Open Multilingual Wordnet. n.d. "Overview." Accessed March 18, 2026. <https://omwn.org/>.
- OpenAI. 2023. "GPT-4 Technical Report." *arXiv Preprint arXiv:2303.08774*. <https://arxiv.org/abs/2303.08774>.
- Orkphol, Korawit, and Wu Yang. 2019. "Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet." *Future Internet* 11 (5): 5. <https://doi.org/10.3390/fi11050114>.
- Pasini, Tommaso, Alessandro Raganato, and Roberto Navigli. 2021. "XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation." *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 13648–56. <https://ojs.aaai.org/index.php/AAAI/article/view/17609>.
- Paumier, Sébastien, Takuya Nakamura, and Stavroula Voyatzi. 2009. "UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources." *eLEX2009: eLexicography in the 21st Century: New Challenges, New Applications. Book of Abstracts*, 173–75. https://www.uclouvain.be/en/system/files/uclouvain_assetmanager/groups/cms-editors-cecl/Drupal%2010%20CECL%20papers/eLEX_book_abstracts_FINAL_New.pdf.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. <https://doi.org/10.3115/v1/D14-1162>.
- Petalinkar, Saša. 2025. "Creating a Synthetic Evaluation Dataset for Serbian SentiWordNet Using Large Language Models." *Infotheca - Journal for Digital Humanities* 24 (1): 53–70. <https://doi.org/10.18485/infotheca.2024.24.1.3>.
- Petalinkar, Saša, Ranka Stanković, Milica Ikonić Nešić, and Cvetana Krstev. 2026. "A Semi-Automated LLM-Based Framework for Word Sense Disambiguation in Serbian." *Intelligent Data Analysis: An International Journal*, August 22, 1088467X261469292. <https://doi.org/10.1177/1088467X261469292>.
- Petalinkar, Saša, Ranka M. Stanković, and Milica Ikonić Nešić. 2025. "Comparative Analysis of Methods for Creating a Sentiment Lexicon of the Serbian WordNet." *The Electronic Library* 43 (4): 547–77. <https://doi.org/10.1108/EL-08-2024-0253>.
- Peters, Matthew, Mark Neumann, Mohit Iyyer, et al. 2018. "Deep Contextualized Word Representations." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–37. <https://doi.org/10.18653/v1/N18-1202>.

- Piper, Predrag, Rajna Dragićević, and Ivan Klajn. 2022. *Normativna gramatika srpskoga jezika*. Četvrto, Izmenjeno i dopunjeno izdanje. Matica srpska.
- “Princeton WordNet.” n.d. Accessed November 30, 2023. <https://wordnet.princeton.edu/>.
- Pu, X., N. Pappas, J. Henderson, and A. Popescu-Belis. 2018. “Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation.” *Transactions of the Association for Computational Linguistics* 6: 635–49. https://doi.org/10.1162/tacl_a_00242.
- Rademaker, Alexandre, Abhishek Basu, and Rajkiran Veluri. 2023. “Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus.” *Proceedings of the 12th Global Wordnet Conference*, 243–53. <https://doi.org/10.18653/v1/2023.gwc-1.30>.
- Raganato, Alessandro, Yves Scherrer, and Jörg Tiedemann. 2019. “The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation.” In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, edited by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, et al. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-5354>.
- Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–90. <https://doi.org/10.18653/v1/D19-1410>.
- Reynolds, Laria, and Kyle McDonell. 2021. “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.” *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, May 8, 1–7. <https://doi.org/10.1145/3411763.3451760>.
- Rios, Annette, Mathias Müller, and Rico Sennrich. 2018. “The Word Sense Disambiguation Test Suite at WMT18.” In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, edited by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, et al. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6437>.
- Rios Gonzales, Annette, Laura Mascarell, and Rico Sennrich. 2017. “Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.” In *Proceedings of the Second Conference on Machine Translation*, edited by Ondřej Bojar, Christian Buck, Rajen Chatterjee, et al. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-4702>.
- Rychlý, Pavel. 2007. “Manatee/Bonito: A Modular Corpus Manager.” In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, edited by Petr Sojka and Aleš Horák. Masaryk University. <https://nlp.fi.muni.cz/raslan/2007/papers/12.pdf>.
- Santoro, Daniela, Beatrice Marchesi, Silvia Zampetta, et al. 2025. “Exploring Latin WordNet Synset Annotation with LLMs.” In *Proceedings of the 13th Global Wordnet Conference*, edited by Chiara Zanchi, Luca Brigada Villa, Erica Biagetti, Alexandre Rademaker, Francis Bond, and German Rigau. Global Wordnet Association. <https://doi.org/10.18653/v1/2025.gwc-1.8>.
- Scherrer, Yves, Alessandro Raganato, and Jörg Tiedemann. 2020. “The MUCOW Word Sense Disambiguation Test Suite at WMT 2020.” In *Proceedings of the Fifth Conference on Machine Translation*, edited by Loïc Barrault, Ondřej Bojar, Fethi Bougares, et al. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.wmt-1.40>.

- Schulhoff, Sander, Michael Ilie, Nishant Balepur, et al. 2024. "The Prompt Report: A Systematic Survey of Prompt Engineering Techniques." Version 6. Preprint, arXiv. <https://doi.org/10.48550/ARXIV.2406.06608>.
- Škorić, Mihailo. 2025. "New Language Models for Serbian." *Infotheca - Journal for Digital Humanities* 24 (1): 7–28. <https://doi.org/10.18485/infotheca.2024.24.1.1>.
- Škorić, Mihailo, and Saša Petalinkar. 2024. "New XLM-R-Based Language Models for Serbian and Serbo-Croatian." *Artificial Intelligence Conference* (Belgrade), 59–60.
- Škorić, Mihailo, and Saša Petalinkar. 2025. *Quality Textual Corpora and New South Slavic Language Models*. <https://doi.org/10.18485/judig.2025.1.ch19>.
- Srivastava, Sumit, Suzanna D. Wentzel, Alejandro Catala, and Mariët Theune. 2025. "Measuring and Implementing Lexical Alignment: A Systematic Literature Review." *Computer Speech & Language* 90: 101731. <https://doi.org/10.1016/j.csl.2024.101731>.
- Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, edited by Ilan Kernerman, Sara Carvalho, Carlos A. Iglesias, and Rachele Sprugnoli. European Language Resources Association. <https://aclanthology.org/2022.salld-1.6>.
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries—from File System to Lemon Based Lexical Database." *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*, 48–56. http://lrec-conf.org/workshops/lrec2018/W23/pdf/book_of_proceedings.pdf.
- Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. "Resource-Based WordNet Augmentation and Enrichment." *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)* (Sofia, Bulgaria), 104–14. <https://aclanthology.org/2018.clib-1.14/>.
- Stanković, Ranka, Rada Stijović, Mihailo Škorić, and Cvetana Krstev. 2025. "The Dictionary of Contemporary Serbian Language (RSSJ): Advanced Automation and Other Challenges." In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, edited by Iztok Kosem, Miloš Jakubiček, Marek Medved', et al. Lexical Computing CZ s.r.o. https://elex.link/elex2025/wp-content/uploads/eLex2025-05-Stankovic_etal.pdf.
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.
- Sumanathilaka, Deshan Koshala, Nicholas Micallef, and Julian Hough. 2024. "Can LLMs Assist with Ambiguity? A Quantitative Evaluation of Various Large Language Models on Word Sense Disambiguation." *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security* (Lancaster, UK), July, 97–108. <https://aclanthology.org/2024.nlpaics-1.12/>.
- Sumanathilaka, Deshan Koshala, Nicholas Micallef, and Julian Hough. 2025. "Prompt Balance Matters: Understanding How Imbalanced Few-Shot Learning Affects Multilingual Sense Disambiguation in LLMs." In *Proceedings of the Workshop on Beyond English: Natural Language Processing for All Languages in an Era of Large Language Models*, edited by

Sudhansu Bala Das, Pruthwik Mishra, Alok Singh, Shamsuddeen Hassan Muhammad, Asif Ekbal, and Uday Kumar Das. INCOMA Ltd., Shoumen, BULGARIA. <https://aclanthology.org/2025.globalnlp-1.2/>.

Sun, Yicheng, and Jie Wang. 2025. “Expanding WordNet Based on Glosses: Methodology and Applications.” *Proceedings of the 13th Global Wordnet Conference*, 37–45.

Tiedemann, Jörg, and Santhosh Thottingal. 2020. “OPUS-MT – Building Open Translation Services for the World.” In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, edited by André Martins, Helena Moniz, Sara Fumega, et al. European Association for Machine Translation. <https://aclanthology.org/2020.eamt-1.61/>.

Tufis, D., D. Cristea, and S. Stamou. 2004. “BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.” *Romanian Journal of Information Science and Technology* 7 (1–2): 9–43.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, et al., vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. “Word-Sense Disambiguation for Machine Translation.” In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, edited by Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Association for Computational Linguistics. <https://aclanthology.org/H05-1097/>.

Vij, Sonakshi, Janhvi Juyal, Amita Jain, and Devendra Tayal. 2025. “Exploring WordNet® Graphs for Text Summarization and Sentiment Analysis in Bengali Speech.” *International Journal of Information Technology* 17 (1): 529–38.

Vitas, Duško, Ranka Stanković, and Cvetana Krstev. 2025. “О фамилији корпуса савременог српског језика СрпКор.” In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*, edited by Jasmina Moskovljević Popović and Ranka Stanković, vol. 1. University of Belgrade — Faculty of Philology. <https://doi.org/10.18485/judig.2025.1.ch1>.

Vossen, Piek. 2004. “EuroWordNet: A Multilingual Database of Autonomous and Language-Specific WordNets Connected via an Inter-Lingual Index.” *International Journal of Lexicography* 17 (2): 161–73. <https://doi.org/10.1093/ijl/17.2.161>.

Vrandečić, Denny, Lydia Pintscher, and Markus Krötzsch. 2023. “Wikidata: The Making Of.” *Companion Proceedings of the ACM Web Conference 2023*, 615–24.

Vreš, Domen, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. “Generative Model for Less-Resourced Language with 1 Billion Parameters.” <https://arxiv.org/abs/2410.06898>.

Vu, Xuan-Son, and Seong-Bae Park. 2014. “Construction of Vietnamese SentiWordNet by Using Vietnamese Dictionary.” *ArXiv abs/1412.8010*.

- Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. “Multilingual E5 Text Embeddings: A Technical Report.” arXiv:2402.05672. Preprint, arXiv, February 8. <https://doi.org/10.48550/arXiv.2402.05672>.
- Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.” Version 2. Preprint, arXiv. <https://doi.org/10.48550/ARXIV.2002.10957>.
- Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, et al. 2023. “Self-Instruct: Aligning Language Models with Self-Generated Instructions.” In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.754>.
- Wankhade, Mayuri, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. “A Survey on Sentiment Analysis Methods, Applications, and Challenges.” *Artificial Intelligence Review* 55: 5731–80.
- Weaver, Warren. 1949. “Translation.” In *Machine Translation of Languages: Fourteen Essays*, edited by William N. Locke and A. Donald Booth. MIT Press.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35: 24824–37.
- Wiegrefe, Sarah, and Ana Marasovic. 2021. “Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing.” In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, edited by Joaquin Vanschoren and Sai-Kit Yeung.
- Wijayanti, Rini, and Andria Arisal. 2021. “Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis.” *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 20: 1–16. <https://doi.org/10.1145/3425632>.
- Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.” In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, edited by Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Association for Computational Linguistics. <https://aclanthology.org/H05-1044/>.
- Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. “Recent Trends in Deep Learning Based Natural Language Processing.” *IEEE Computational Intelligence Magazine* 13 (3): 55–75. <https://doi.org/10.1109/MCI.2018.2840738>.
- Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. “Sentiment Analysis in the Era of Large Language Models: A Reality Check.” *Findings of the Association for Computational Linguistics: NAACL 2024*, 3881–906. <https://doi.org/10.18653/v1/2024.findings-naacl.246>.

9 Прилог термиолошки глосар

У овом термиолошком глосару издвојени су појмови који су важни за разумевање теоријског, методолошког и техничког дела дисертације. За сваки термин наводи се српски назив, енглески еквивалент и кратко објашњење његовог значења у контексту овог рада. Енглески еквиваленти су задржани ради једнозначног повезивања са стручном литературом из области обраде природног језика, машинског учења и лексичко-семантичких ресурса, док су дефиниције формулисане тако да подрже доследну употребу терминологије у дисертацији.

аблација промптова — prompt ablation

Провера у којој се из промпта уклања или се у њему мења једна његова компонента како би се проценио њен утицај на резултат модела.

алгоритам насумичних шетњи — random walk

Стохастички поступак кретања по чворовима графа, при чему се у сваком кораку избор наредног чвора врши насумично, у складу са дистрибуцијом вероватноће дефинисаном над излазним гранама текућег чвора.

алгоритам потпорних вектора — Support Vector Machine (SVM)

Метод надгледаног учења заснован на конструкцији хиперравни која максимализује маргину између класа у векторском простору обележја.

анализа сентимента — sentiment analysis

Задатак аутоматског препознавања и класификације текстуалних исказа према израженом сентименту, који се најчешће одређује као позитиван, негативан или неутралан.

анотатор — annotator

Особа или систем који додељује ознаке текстуалним јединицама у поступку аотације.

анотација — annotation

Поступак додељивања лингвистичких, семантичких или других ознака текстуалним јединицама, односно сегментима текста који се састоје од једног или више токена.

анотација сентимента — sentiment annotation

Аотација којом се текстуалној јединици, лексеми или синсету придружује информација о сентименту.

анотирани сегмент текста — annotated span

Континуални део (сегмент) текста, на пример једна реч или вишечлана јединица, коме је у корпусу придружена аотација.

антонимија — antonymy

Лексичка или семантичка релација супротности између две лексеме или два значења.

аргмакс — argmax

Операција избора елемента или индекса који има највећу вредност задате функције или мере, на пример, највећу косинусну сличност.

$$\operatorname{argmax}_x f(x)$$

архитектура мешавине експерата — mixture-of-experts architecture

Архитектура у којој се при обради појединачног улаза активира само подскуп специјализованих компоненти, односно експерата, чиме се повећава укупан капацитет модела без пропорционалног повећања активног рачунарског трошка.

аутоматска анотација — automatic annotation

Поступак у коме систем без непосредне људске интервенције додељује ознаке текстуалним јединицама.

базни модел — foundation model

Велики претходно обучен модел, обучен на општим, обимним скуповима података, који служи као основа за даљу примену, укључујући фино подешавање или употребу у режимима без додатне обуке.

Бернулијев наивни Бајесов класификатор — Bernoulli Naive Bayes classifier

Варијанта наивног Бајесовог класификатора намењена бинарним обележјима, користи се као један од класичних модела у евалуацији.

векторска репрезентација — embedding

Густо нумеричко представљање лексичке јединице, документа или другог ентитета у облику вектора у континуалном простору.

векторска репрезентација речи — word embedding

Нумерички вектор који представља реч у континуалном простору тако да се семантичка или дистрибутивна сличност одражава у геометријској блискости вектора.

велики језички модели (BJM) — large language models (LLMs)

Дубоки модели засновани на неуронским мрежама, најчешће засновани на трансформерској архитектури, обучени на великим количинама текстуалних података и способни за широк спектар задатака обраде природног језика.

Вилсонов интервал поверења — Wilson confidence interval

Интервал поверења за пропорцију који се често користи када се процењује тачност или други удео на ограниченом броју примера.

вишеглави механизам пажње — multi-head attention

Варијанта механизма пажње у којој се пажња рачуна кроз више паралелних пројекција, чиме модел истовремено захвата различите типове односа између елемената секвенце.

вишезначност — lexical ambiguity; polysemy

Својство лексеме да има више различитих значења; у ужем смислу може се односити на полисемију, а у ширем може обухватати и друге облике вишезначности.

вишечлана јединица — multiword expression

Лексичка јединица састављена од две или више речи која у одређеном контексту има јединствено значење које се најчешће не може у потпуности извести из значења појединачних компоненти.

ворднет — WordNet

Лексичко-семантичка мрежа заснована на организацији лексемау синсете и њиховом повезивању семантичким релацијама.

генеративни модел — generative model

Модел који на основу улазног контекста производи нови излаз, најчешће текст, ознаку, дефиницију или структурирани одговор (овде се ограничавамо на текст који је предмет ове дисертације).

гласа — gloss

Кратак дефиницијски опис значења синсета у ворднету. Гласа у ворднету не садржи примере употребе, иако их у неким другим ресурсима може обухватати.

декодер — decoder

Део трансформерске архитектуре или модел који ауторегресивно генерише излазни низ токена на основу претходног контекста.

дестилација модела — model distillation

Поступак преноса знања са већег или сложенијег модела на мањи модел, обично ради ефикасније примене уз очување дела перформанси.

дуга краткорочна меморија (LSTM) — Long Short-Term Memory

Варијанта архитектуре рекурентне неуронске мреже осмишљена за ефикасније моделирање дугорочних зависности у секвенцијалним подацима.

евалуациона јединица — evaluation unit

Основна јединица над којом се рачуна или процењује резултат евалуације, на пример реченица, токен, лексема, синсет или документ.

евалуациони скуп / скуп за евалуацију — evaluation dataset

Скуп података који се користи за мерење перформанси модела, метода или ресурса.

експертска валидација — human validation

Поступак у коме човек проверава и потврђује исправност аутоматски добијених резултата.

екстерна евалуација — external evaluation

Провера ресурса или модела на задатку који није непосредно коришћен при његовој изградњи, како би се проценила практична употребљивост добијеног резултата.

енкодер — encoder

Део трансформерске архитектуре или модел који улазни текст претвара у контекстуалне репрезентације погодне за класификацију, претраживање, поређење или друге низводне задатке.

закључивање, генерисање — inference

Фаза у којој се већ обучени модел примењује на нове податке ради генерисања предвиђања или излазних резултата без даљег ажурирања параметара.

затворени комерцијални модел — closed commercial model

Модел који је доступан као комерцијални сервис или програмски интерфејс, без јавног приступа обученим тежинама, подацима за обуку или потпуним техничким детаљима архитектуре.

златни скуп — gold dataset; gold standard

Ручно проверен и поуздано означен скуп података који се користи као референтна основа за евалуацију модела, метода или ресурса.

идентификатор синсета — synset identifier

Јединствена ознака која служи за недвосмислено упућивање на одређени синсет у ворднету или другом лексичко-семантичком ресурсу.

издвојени скуп за тестирање — held-out test set

Део података који се у датој итерацији не користи за обуку, подешавање модела, упита или ревизију ресурса, већ служи за процену текуће верзије поступка.

излазни податак — output

Резултат који модел или алгоритам производи након обраде улазних података.

инвентар значења — sense inventory

Унапред дефинисан скуп могућих значења лексеме из којег се у задатку разрешења вишезначности бира значење остварено у датом контексту.

инжењерство промптова — prompt engineering

Поступак систематског формулисања, структурирања и оптимизације инструкција (промптова) са циљем постизања прецизнијих или контролисанијих излазних резултата модела.

интервал поверења — confidence interval

Опсег вредности који, под задатим статистичким претпоставкама, описује неизвесност процене посматране мере.

$$IP = \hat{\theta} \pm c \cdot SE$$

кандидатско значење / кандидат за значење — candidate sense

Једно од могућих значења које се разматра као потенцијално одговарајуће за циљну реч у датом контексту.

картица модела — model card

Документ који описује намену, техничке особине, податке за обуку, ограничења, лиценцу и препоручени начин употребе модела.

категорија исхода — outcome category

Класа која описује исход поређења двеју или више поставки, на пример, да ли је информација о значењу побољшала превод циљне јединице, оставила га непромењеним или га погоршала.

Кендалов τ коефицијент рангиране корелације — Kendall's tau rank correlation coefficient

Непараметарска мера статистичке сагласности између два ранга. Заснива се на поређењу свих могућих парова елемената и дефинише се као нормализована разлика између броја сагласних и несагласних парова. За две рангиране листе са n елемената, Кендалов τ се рачуна као:

$$\tau = \frac{C - D}{\binom{n}{2}}$$

класификатор — classifier

Модел машинског учења који улазним примерима придружује једну од унапред дефинисаних класа.

контекст — context

Језичко окружење у коме се јавља циљна реч и које се користи за тумачење њеног значења.

контекстни прозор — context window

Највећи број токена који модел може да прими и обради у једном улазу, укључујући промпт додатни контекст и, у појединим поставкама, претходни ток разговора.

контекстуална репрезентација — contextual representation

Репрезентација лексеме или другог текстуалног елемента која зависи од конкретног језичког контекста у коме се та јединица појављује.

контролни узорак — control sample

Ограничен и унапред дефинисан део података који се користи за проверу, поређење или праћење резултата у више фаза истраживања.

косинусна сличност — cosine similarity

Мера сличности између два вектора заснована на косинусу угла између њих, често коришћена за поређење векторских репрезентација.

$$\cos(x,y) = (x \cdot y) / (||x|| \cdot ||y||)$$

лексема — lexical unit

Јединица лексичког описа која може бити представљена једном речју или вишечланим изразом и која се у ресурсу третира као носилац једног или више значења.

лексикон сентимента — sentiment lexicon

Лексички ресурс у коме су лексичким јединицама или значењима придружене вредности поларитета сентимента.

лексички вишезначна јединица — lexically ambiguous unit

Лексема која у датом језику има више могућих значења, због чега њено тумачење или превођење зависи од контекста.

лексичко-семантичко поравнање — lexical-semantic alignment

Шири поступак успостављања веза између лексичких јединица, значења, дефиниција или синсета у различитим лексичко-семантичким ресурсима.

лема — lemma

Основни, канонски облик лексеме под којим су обједињени њени флективни облици и који може бити повезан са једним или више значења.

лематизација — lemmatization

Поступак придруживања канонског облика тј. леме флективном облику речи (у тексту).

литерал — literal

Лексичка форма која припада конкретном синсету, односно појава леме у оквиру одређеног значења.

Макнимаров тест — McNemar's test

Статистички тест за поређење два повезана класификациона поступка на истим примерима, заснован на броју случајева у којима се поступци разликују.

$$\chi^2 = (|b - c| - 1)^2 / (b + c)$$

макро-F1 мера — macro-F1 score

Просек F1-мера израчунатих по класама, при чему свака класа има једнаку тежину без обзира на њену учесталост у скупу података.

$$\text{макро-F1} = (F1_1 + \dots + F1_k) / K$$

мапирање синсета — synset mapping

Поступак успостављања везе између синсета у различитим ресурсима или различитим верзијама истог ресурса.

машинско превођење — machine translation

Задатак аутоматског превођења текста са једног природног језика на други помоћу рачунарског система или језичког модела.

међујезички индекс — Inter-Lingual Index (ILI)

Посреднички индекс који омогућава повезивање синсета различитих језика преко заједничког концептуалног идентификатора.

меронимија — meronymy

Семантичка релација део-целина у којој један појам, мероним, означава део другог појма, холонима.

механизам пажње — attention mechanism

Механизам у моделима заснованим на неуронским мрежама који одређује релативну релевантност различитих делова улазне секвенце при формирању репрезентације или излаза модела.

механизам самопажње — self-attention

Облик механизма пажње у коме се репрезентација сваког елемента у секвенци израчунава у односу на све остале елементе исте секвенце.

модел вреће токена — bag-of-tokens model

Модел репрезентације текста у коме се документ описује вектором учесталости појединих токена, при чему се занемарују њихов редослед и структурне зависности.

модел дообучен инструкцијама — instruction-tuned model

Модел који је након претходне обуке додатно прилагођен да следи упутства на природном језику и да производи одговоре у облику погодном за разговорну употребу или употребу усмерену према задацима.

модел са отвореним тежинама — open-weight model

Модел чије су обучене тежине јавно доступне за преузимање, локално покретање или даље прилагођавање, у оквиру услова одговарајуће лиценце.

моноксемска јединица — single-word expression

Лексема представљена једном речју.

морфосинтаксичко обележавање — morphosyntactic tagging

Поступак додељивања морфолошких и синтаксичких ознака токенима у тексту, као што су врста речи, род, број, падеж или глаголски облик.

надгледано учење — supervised learning

Методологија машинског учења у којој је сваком улазном примеру придружена позната излазна ознака.

наивни Бајесов класификатор — Naive Bayes classifier (NB)

Метод надгледаног учења заснован на Бајесовој теореме и претпоставци условне независности обележја дате класе.

напомена о значењу / семантичка смерница — sense hint; sense note

Додатна информација о значењу циљне лексеме која се доставља моделу како би се усмерио избор одговарајућег тумачења или преводног еквивалента.

недодељени исход — unassigned outcome; unmatched outcome

Исход евалуације у коме поступак није издвојио ниједно од понуђених значења или није успоставио очекивано повезивање.

ненадгледано учење — unsupervised learning

Методологија машинског учења у којој излазне ознаке нису доступне током обучавања.

низводни задатак — downstream task

Конкретан задатак примене модела или ресурса, као што су класификација текста, анализа сентимента или разрешење вишезначности речи.

обележје — feature

Квантитативна или категоријална карактеристика која се користи као улазна променљива у моделу машинског учења.

обогаћивање ресурса — resource enrichment

Поступак допуњавања постојећег језичког или лексичко-семантичког ресурса новим јединицама, значењима, дефиницијама, примерима или ознакама.

обрада природног језика (ОПЈ) — natural language processing (NLP)

Истраживачка и примењена област рачунарства која се бави развојем формалних модела и алгоритама за аутоматску анализу, репрезентацију и обраду података на природном језику.

обука — training

Фаза у машинском учењу у којој се параметри модела прилагођавају на основу скупа података ради минимизације функције губитка.

одговор — response

Излазни текст или предвиђање које модел генерише као резултат обраде датог промпта.

одзив — recall

Однос броја позитивних примера који су тачно класификовани и укупног броја стварно позитивних примера.

$$R = TP / (TP + FN)$$

омотач — wrapper

Софтверски модул који обавија постојећу функционалну компоненту или библиотеку и обезбеђује додатни интерфејс, адаптацију или контролу приступа без измене њене унутрашње логике.

оријентација сентимента / усмереност сентимента — sentiment orientation

Вредносна усмереност исказа или лексеме која може бити позитивна, негативна или неутрална.

оцена сентимента — sentiment score

Нумеричка вредност којом се лексичкој јединици, синсету или другом елементу придружује степен позитивног, негативног или објективног сентимента.

паралелни корпус — parallel corpus

Корпус који садржи текстове или реченице на два или више језика и који су међусобно поравнати као преводни еквиваленти.

позиционо кодирање — positional encoding

Механизам у трансформерској архитектури којим се векторским репрезентацијама токена додаје информација о њиховом положају у секвенци.

покривеност значења — sense coverage

Степен у коме одређени ресурс садржи значења потребна за обраду, анотацију или евалуацију датог језичког материјала.

покривеност инвентара — inventory coverage

Степен у коме инвентар значења садржи значења потребна за анотацију, разрешење вишезначности или евалуацију посматраног корпуса.

полазна поставка — baseline setting; baseline

Референтна поставка, модел или верзија ресурса са којом се пореде нови поступци или обогаћене верзије.

поларитет сентимента — sentiment polarity

Операциона репрезентација оријентације сентимента у облику дискретне класе или нумеричке вредности.

полилексемска јединица, вишечлана јединица — multiword lexical unit

Лексема састављена од више речи које заједно представљају једно значење или један појам.

полисемија — polysemy

Облик вишезначности у коме једна лексема има више међусобно повезаних значења.

полуаутоматски поступак — semi-automatic procedure

Поступак у коме се аутоматска обрада комбинује са људском провером, корекцијом или валидацијом.

поравнање значења речи — word sense alignment

Поступак повезивања значења из два или више лексичко-семантичких ресурса када се процени да упућују на исти или довољно близак концепт.

поставка улаза — input configuration

Начин на који су улазни подаци организовани и достављени моделу, укључујући реченицу, додатне напомене, ограничења и друге информације релевантне за задатак.

поткорпус — subcorpus

Део већег корпуса издвојен према језику, домену, жанру, анотационом слоју или другом релевантном критеријуму.

праг одлучивања — decision threshold

Вредност на основу које се одређује да ли ће одређено предвиђање, оцена или мера бити прихваћена као довољна за доделу ознаке или доношење одлуке.

преводни еквивалент — translation equivalent

Лексема или израз у циљном језику који у одређеном контексту одговара лексеми или значењу из изворног језика.

претходна обука — pretraining

Фаза у којој се модел обучава на великој количини података ради стицања општих репрезентација и параметара који се касније могу прилагођавати конкретном задатку.

прецизност — precision

Однос броја позитивних примера који су тачно класификовани и укупног броја позитивно класификованих примера.

$$P = TP / (TP + FP)$$

пример употребе — usage example

Реченица или кратак исказ који илуструје употребу лексеме или синсета у конкретном језичком контексту.

Принстонски ворднет (PWN) — Princeton WordNet

Референтна лексичко-семантичка мрежа за енглески језик, развијена на Универзитету Принстон; организује лексеме у синсете и повезује их семантичким и лексичким релацијама. У овом раду се често помиње и користи као извор за поравнање и као полазни ресурс за вишејезичке и националне ворднете.

програмски интерфејс (API) — application programming interface (API)

Скуп правила и функција преко којих програм приступа другом систему, библиотеци, сервису или моделу.

промпт — prompt

Текстуална инструкција или улазни садржај који се доставља моделу ради иницирања генерисања одговора.

развојно-аналитички скуп — development/analysis set

Ручно проверен део података који се користи за анализу грешака, ревизију ресурса и развој поступка, а не као независна завршна процена.

разрешење вишезначности речи — Word Sense Disambiguation (WSD)

Задатак аутоматског одређивања исправног значења полисемичне лексеме у конкретном контексту.

рангирање значења кандидата — candidate sense ranking

Поступак у коме се кандидатска значења ређају према процењеној прикладности за дату лексему у контексту.

рандомизациони тест — randomization test

Непараметарски тест у коме се статистичка значајност процењује на основу расподеле добијене насумичним премештањем или заменом исхода.

редакција ресурса — resource revision

Поступак систематске провере и исправке постојећих елемената језичког или лексичко-семантичког ресурса.

режим без примера — zero-shot

Начин примене модела у коме се задатак задаје без иједног наведеног примера решења, ослањајући се искључиво на инструкцију и претходно стечено знање модела.

режим са малим бројем примера — few-shot

Начин примене модела у коме се задатак задаје уз мали број репрезентативних примера који моделу служе као образац за формирање одговора.

рекурентне неуронске мреже (RNN) — recurrent neural networks

Класа неуронских модела у којима се улазна секвенца обрађује редом, уз преношење информације кроз низ узастопних стања.

репрезентација — representation

Формални начин представљања података или лингвистичких јединица у моделу, који може бити симболички, векторски или хибридни.

референтна ознака — reference label

Ручно проверена или методолошки прихваћена ознака која служи као основа за поређење са аутоматски добијеним излазом.

референтни превод — reference translation

Превод који се користи као основа за поређење и евалуацију аутоматски добијеног превода.

ручна евалуација — manual evaluation

Поступак у коме човек процењује тачност, прикладност или квалитет резултата добијених аутоматском обрадом.

самодорада — self-refinement

Поступак у коме модел накнадно процењује, преиспитује и по потреби коригује сопствени излаз ради повећања тачности, стабилности или усклађености са задатим ограничењима.

семантичка грануларност — semantic granularity

Ниво детаљности са којим су значења разграничена у лексичко-семантичком ресурсу.

семантичка мрежа — semantic network

Лексичко-семантички ресурс у коме су концепти (значења) организовани у структуру повезану формално дефинисаним семантичким релацијама.

семантичка релација — semantic relation

Формално дефинисан однос између значења или синсета, као што су хиперонимија, хипонимија, меронимија, холонимија или антонимија.

синсет — synset

Скуп лексема које деле исто значење и представљају један концепт у оквиру структуре ворднета.

синтетички скуп за евалуацију — synthetic evaluation dataset

Евалуациони скуп чији су примери добијени аутоматским генерисањем, а затим коришћени за проверу модела или ресурса у контролисаним условима.

слој за разрешење вишезначности речи — WSD layer

Анотациони слој или слој добијен као резултат обраде у коме су појавама лексема у тексту придружена одговарајућа значења из инвентара значења.

сребрни скуп — silver dataset

Аутоматски или полуаутоматски означен скуп података који се користи за обуку или проширење податакакоји има нижи степен поузданости од златног скупа.

Српски ворднет (SrpWN) — Serbian WordNet

Лексичко-семантичка мрежа за српски језик развијена у оквиру EuroWordNet/BalkaNet традиције, у којој су значења представљена синсетима и повезана семантичким релацијама; ресурс је поравнат са Принстонским ворднетом посредством међујезичког индекса (ILI).

стоп-речи — stop words

Фреквентне функционалне речи које се у појединим поступцима ОПЈ уклањају ради редукције димензионалности репрезентације текста.

структурирани излаз — structured output

Излаз модела организован у унапред дефинисана поља или формат, тако да се може програмски проверити и обрадити.

тачност — accuracy

Мера учинка дефинисана као однос броја тачно класификованих или исправно означених примера и укупног броја примера.

$$T = (TP + TN) / (TP + TN + FP + FN)$$

токен — token

Елементарна јединица текста добијена поступком сегментације, која може представљати реч, карактер, н-грам или други низ симбола.

токенизација — tokenization

Поступак сегментације текста на токене, односно елементарне јединице погодне за даљу рачунарску обраду.

трансформерска архитектура — transformer architecture

Архитектура дубоког учења заснована на механизмима пажње, у којој се међусобни односи елемената у улазној секвенци моделирају без рекурентне обраде.

улазни податак — input

Податак, текст или структура који се достављају моделу или алгоритму ради обраде.

уланчавање промпта — prompt chaining

Поступак разлагања сложеног задатка на низ међусобно повезаних промптова, при чему излаз претходног корака служи као улаз, контекст или ограничење у наредном кораку.

универзална врста речи — universal part of speech (UPOS)

Општа категорија врсте речи дефинисана у оквиру универзалних схема морфосинтаксичког обележавања.

учење преносом знања — transfer learning

Поступак у коме се параметри или репрезентације модела обученог на изворном домену користе као почетна основа за обуку на циљном домену.

учење у контексту — in-context learning

Способност модела да на основу инструкције и примера датих у самом улазу прилагоди понашање у задатку без ажурирања својих параметара.

фино подешавање — fine-tuning

Поступак накнадног прилагођавања претходно обученог модела конкретном задатку или домену на основу циљног скупа података.

фино подешени модел — fine-tuned model

Модел који је након претходне обуке додатно прилагођен конкретном задатку или домену поступком финог подешавања.

Флајсова κ мера — Fleiss' kappa

Непараметарска мера сагласности између више анотатора за категоријалне податке, дефинисана као нормализована разлика између посматране и очекиване случајне сагласности.

$$\kappa = (\bar{P} - P_e) / (1 - P_e)$$

формат промпт–одговор — prompt–response format

Структурна поставка интеракције са моделом у којој се задатак представља путем промпта, а модел генерише одговор као излаз.

Фридманов тест — Friedman test

Непараметарски тест за анализу варијансе поновљених мерења који се користи за утврђивање постојања статистички значајних разлика између више зависних група.

халуцинација модела — hallucination

Појава у којој генеративни модел производи садржај који је формално уверљив, али фактички нетачан, непоткрепљен или измишљен.

хиперонимија — hypernymy

Семантичка релација надређености у којој један појам представља општију категорију у односу на други појам.

хипонимија — hyponymy

Семантичка релација подређености у којој један појам представља ужу категорију у односу на надређени појам.

Холмова корекција — Holm correction

Поступак корекције р-вредности при више истовремених статистичких поређења, којим се смањује ризик од лажно позитивних закључака.

холонимија — holonymy

Семантичка релација целина–део у којој један појам, холоним, означава целину која обухвата други појам, мероним, као свој део.

хомонимија — homonymy

Појава да исти лексички облик има два или више међусобно неповезаних значења.

циљана евалуација — targeted evaluation

Евалуација усмерена на унапред одређену појаву, подскуп примера или специфичан аспект понашања модела, уместо на општу успешност система у целини.

циљна лексема — target lexical unit

Лексема у тексту која је предмет анализе, разрешења вишезначности, анотације или провере преводног еквивалента.

циљна реч — target word

Реч или лексема у тексту за коју се одређује одговарајуће значење, синсет или друга семантичка ознака.

AdaBoost — Adaptive Boosting

Метод машинског учења који комбинује више слабијих класификатора у јачи модел тако што итеративно повећава тежине погрешно класификованих примера.

F1-мера — F1-score

Хармонијска средина прецизности и одзива.

$$F1 = 2 \cdot P \cdot R / (P + R)$$

JSON формат — JSON format

Текстуални формат за структурирано представљање података, коришћен у овом раду за контролу и проверу излаза модела.

TF-IDF — Term Frequency – Inverse Document Frequency

Мера тежине термина у документу дефинисана као производ учесталости термина у документу и инверзне учесталости тог термина у корпусу:

$$TF\text{-}IDF(t, d) = TF(t, d) \cdot IDF(t)$$

У овој формули $TF(t, d)$ означава учесталост термина t у документу d , а $IDF(t)$ логаритамски скалирану инверзну учесталост тог термина у корпусу.

XML репрезентација, XML серијализација — XML representation

Представљање лексичко-семантичког ресурса у XML формату, погодно за размену, претраживање и програмску обраду.

10 Прилог индекс табела

ТАБЕЛА 2.2 ОСНОВНИ ПОЈМОВИ ЗА ОПИС СТРУКТУРЕ ВОРДНЕТА	9
ТАБЕЛА 3.1 КОНФИГУРАЦИЈЕ ЗА ЕКСТЕРНУ ЕВАЛУАЦИЈУ У АНАЛИЗИ СЕНТИМЕНТА	58
ТАБЕЛА 4.1 СТАТИСТИКА АНОТИРАНИХ ЗНАЧЕЊА ПО ВРСТАМА РЕЧИ И ТИПУ ЈЕДИНИЦЕ; КОРПУС ОБУХВАТА 2.024 РЕЧЕНИЦЕ.	63
ТАБЕЛА 4.2 ПРЕГЛЕД КОРИШЋЕНИХ ВЕЛИКИХ ЈЕЗИЧКИХ И РЕПРЕЗЕНТАЦИОНИХ МОДЕЛА	67
ТАБЕЛА 5.1 МЕТРИКЕ КЛАСИФИКАЦИЈЕ ДОБИЈЕНЕ НА ЕВАЛУАЦИОНОМ СКУПУ ЗА МОДЕЛЕ ADABOOST, NB, SVM, RNN, TRANSFORMER, BERT ¹ , GPT2-ORAO И JERTEN-355	72
ТАБЕЛА 5.2 РЕЗУЛТАТИ ЕВАЛУАЦИЈЕ КОНСТРУИСАНИХ ЛЕКСИКОНА СЕНТИМЕНТА И ЊИХОВО ПОРЕЂЕЊЕ СА ИНИЦИЈАЛНИМ ЛЕКСИКОНОМ S0	74
ТАБЕЛА 5.3 РАСПОДЕЛА ВРЕДНОСТИ ПОЗИТИВНЕ И НЕГАТИВНЕ КОМПОНЕНТЕ У УРАВНОТЕЖЕНОМ УЗОРКУ СИНСЕТА	75
ТАБЕЛА 5.4 САГЛАСНОСТ АНОТАТОРА У РУЧНОЈ ПРОВЕРИ СИНТЕТИЧКОГ СКУПА ЗА ЕВАЛУАЦИЈУ	76
ТАБЕЛА 5.5 ОДНОС ВЕЋИНСКИХ КОНАЧНИХ ВРЕДНОСТИ АНОТАТОРА И ВРЕДНОСТИ ДОБИЈЕНИХ МОДЕЛОМ	77
ТАБЕЛА 5.6 ПОРЕЂЕЊЕ ЛЕКСИКОНА КАНДИДАТА СЕНТИМЕНТА СА СИНТЕТИЧКИМ ЕВАЛУАЦИОНИМ СКУПОМ ИЗ ОДЕЉКА 5.2.3	78
ТАБЕЛА 5.7 ПОРЕЂЕЊЕ БРОЈА ОДРЕДНИЦА, ЗНАЧЕЊА, ЛЕМА И ДРУГИХ ПОКАЗАТЕЉА У ВЕРЗИЈАМА WSD-SR-SENS-INV-V1 И WSD-SR-SENS-INV-V2	84
ТАБЕЛА 5.8 РАЗЛИКЕ КОД ЈЕДНОЧЛАНИХ ОДРЕДНИЦА ПО КАТЕГОРИЈАМА UPOS У ВЕРЗИЈАМА WSD-SR-SENS-INV-V1 И WSD-SR-SENS-INV-V2	84
ТАБЕЛА 5.9 РАЗЛИКЕ КОД ВИШЕЧЛАНИХ ОДРЕДНИЦА ПО КАТЕГОРИЈАМА UPOS У ВЕРЗИЈАМА WSD-SR-SENS-INV-V1 И WSD-SR-SENS-INV-V2	85
ТАБЕЛА 5.10 УЛОГА МОДЕЛА И ВАРИЈАНТИ ПОСТУПКА У ЕВАЛУАЦИЈИ РАЗРЕШЕЊА ВИШЕЗНАЧНОСТИ РЕЧИ	91
ТАБЕЛА 5.11 ТАЧНОСТ МОДЕЛА НА ЈЕДНОЧЛАНИМ ЈЕДИНИЦАМА У РАЗВОЈНО-АНАЛИТИЧКОМ СКУПУ ОД 120 РЕЧЕНИЦА, ГДЕ R1 И R2 ОЗНАЧАВАЈУ ПРВУ И ДРУГУ ИТЕРАЦИЈУ ЕВАЛУАЦИЈЕ	92
ТАБЕЛА 5.12 ТАЧНОСТ МОДЕЛА НА ВИШЕЧЛАНИМ ЈЕДИНИЦАМА У РАЗВОЈНО-АНАЛИТИЧКОМ СКУПУ ОД 120 РЕЧЕНИЦА	93
ТАБЕЛА 5.13 ТАЧНОСТ МОДЕЛА НА ЈЕДНОЧЛАНИМ ЈЕДИНИЦАМА У ИЗДВОЈЕНОМ ТЕСТ СКУПУ ОД 300 РЕЧЕНИЦА	96
ТАБЕЛА 5.14 ТАЧНОСТ МОДЕЛА НА ВИШЕЧЛАНИМ ЈЕДИНИЦАМА У ИЗДВОЈЕНОМ ТЕСТ СКУПУ ОД 300 РЕЧЕНИЦА	97
ТАБЕЛА 5.15 НАЈПРОБЛЕМАТИЧНИЈЕ ЛЕМЕ НА ИЗДВОЈЕНОМ ТЕСТ СКУПУ	102
ТАБЕЛА 5.16 ПОРЕЂЕЊЕ ПОЛАЗНИХ И ДЕСТИЛОВАНИХ МОДЕЛА НА ИЗДВОЈЕНОМ ТЕСТ СКУПУ	104
ТАБЕЛА 5.17 ПАРНА ПРОВЕРА РАЗЛИКЕ ИЗМЕЂУ ПОЛАЗНИХ И ДЕСТИЛОВАНИХ МОДЕЛА	105
ТАБЕЛА 5.18 ПОРЕЂЕЊЕ ДЕСТИЛОВАНИХ МОДЕЛА	105
ТАБЕЛА 5.19 ПРОМЕНА РЕЗУЛТАТА КОД ЈЕДНОЧЛАНИХ ЈЕДИНИЦА	106
ТАБЕЛА 5.20 ПРОМЕНА РЕЗУЛТАТА КОД ВИШЕЧЛАНИХ ЈЕДИНИЦА	106
ТАБЕЛА 5.21 ПРОМЕНА ТАЧНОСТИ ПРЕМА БРОЈУ КАНДИДАТСКИХ ЗНАЧЕЊА: SIMPLE	106
ТАБЕЛА 5.22 ПРОМЕНА ТАЧНОСТИ ПРЕМА БРОЈУ КАНДИДАТСКИХ ЗНАЧЕЊА: MULTILINGUAL-E5-LARGE	107
ТАБЕЛА 5.23 ПРОМЕНА ТАЧНОСТИ ПРЕМА БРОЈУ КАНДИДАТСКИХ ЗНАЧЕЊА: TESLA	107
ТАБЕЛА 5.24 ПОЛОЖАЈ НАЈБОЉЕГ ДЕСТИЛОВАНОГ МОДЕЛА У ОДНОСУ НА ГЕНЕРАТИВНЕ МОДЕЛЕ	108
ТАБЕЛА 5.25 ОБИМ ГЕНЕРИСАНЕ СЕРИЈЕ И РУЧНО ПРОВЕРЕНОГ ПОДСКУПА	109
ТАБЕЛА 5.26 РАСПОДЕЛА ПУНЕ СЕРИЈЕ ПО ВРСТИ РЕЧИ	109
ТАБЕЛА 5.27 РАСПОДЕЛА РУЧНО ПРОВЕРЕНОГ ПОДСКУПА	109
ТАБЕЛА 5.28 ПРОСЕЧНЕ ОЦЕНЕ ПО РАДНОМ ТОКУ	110
ТАБЕЛА 5.29 ПОКАЗАТЕЉИ ИЗ РУЧНЕ ПРОВЕРЕ	111
ТАБЕЛА 5.30 ИЛУСТРАТИВНИ ПРИМЕРИ ИЗ РУЧНЕ ПРОВЕРЕ	112
ТАБЕЛА 5.31 ОСНОВНИ ПОКАЗАТЕЉИ СКУПА ЗА ПОРАВНАЊЕ	114
ТАБЕЛА 5.32 ПОКАЗАТЕЉИ РАЗЛИКА У ГРАНУЛАРНОСТИ РЕЧНИЧКИХ ЗНАЧЕЊА И СИНСЕТА	115
	179

ТАБЕЛА 5.33 РЕЗУЛТАТИ ЕКСТЕРНЕ ЕВАЛУАЦИЈЕ У АНАЛИЗИ СЕНТИМЕНТА РЕЧЕНИЦА	119
ТАБЕЛА 5.34 ТЕСТОВИ СТАТИСТИЧКЕ ЗНАЧАЈНОСТИ ЗА ГЛАВНА ПОРЕЂЕЊА	121
ТАБЕЛА 5.35 ОБИМ ПРИПРЕМЉЕНОГ СКУПА ЗА GOOGLE BASIC ПРОВЕРУ	123
ТАБЕЛА 5.36 КОПИРАЊЕ УМЕТНУТИХ [ENG: ...] ОЗНАКА У GOOGLE BASIC ИЗЛАЗИМА	124
ТАБЕЛА 5.37 ИЗДВАЈАЊЕ ПОДСКУПА ПРИМЕРА ЗАХТЕВНИЈИХ ЗА РЕШАВАЊЕ	124
ТАБЕЛА 5.38 ДИЈАГНОСТИЧКИ ПОКАЗАТЕЉИ ЗА M2M100 НА ПОДСКУПУ ПРИМЕРА ЗАХТЕВНИЈИХ ЗА РЕШАВАЊЕ	125

11 Прилог индекс слика

СЛИКА 3-1 РАДНИ ТОК ФОРМИРАЊА ЛЕКСИКОНА СЕНТИМЕНТА ЗА СРПСКИ ЈЕЗИК НА ОСНОВУ СРПСКОГ ВОРДНЕТА, ПРОШИРИВАЊА СКУПА ПОЛАРИТЕТА И ОБУЧАВАЊА МОДЕЛА МАШИНСКОГ УЧЕЊА.....	35
СЛИКА 3-2 СХЕМАТСКИ ПРИКАЗ РАДНОГ ТОКА РАЗРЕШЕЊА ЗНАЧЕЊА РЕЧИ ПРИМЕНОМ ГЕНЕРАТИВНОГ МОДЕЛА.....	47
СЛИКА 3-3 ИТЕРАТИВНИ ПРОЦЕС АУТОМАТСКЕ АНОТАЦИЈЕ, РУЧНЕ КОРЕКЦИЈЕ И ДОПУНЕ ИНВЕНТАРА ЗНАЧЕЊА.....	48
СЛИКА 3-4 ПРИКАЗ РАДНИХ ТОКОВА ЗА ПРОШИРИВАЊЕ СРПСКОГ ВОРДНЕТА.....	54
СЛИКА 4-1 ПРЕГЛЕД КОРИШЋЕНИХ РЕСУРСА, СКУПОВА ПОДАТАКА И МОДЕЛА У ЕКСПЕРИМЕНТАЛНОМ ОКРУЖЕЊУ	60
СЛИКА 5-1 РЕЗУЛТАТИ РУЧНЕ ПРОВЕРЕ СИНТЕТИЧКОГ СКУПА ЗА ЕВАЛУАЦИЈУ ПО АНОТАТОРУ ...	76
СЛИКА 5-2 ИНТЕРФЕЈС АЛАТА INSERTION.....	83
СЛИКА 5-3 АНОТИРАНИ СЕГМЕНТИ ТЕКСТА ПРЕМА БРОЈУ ПОНУЂЕНИХ МОГУЋИХ ЗНАЧЕЊА У РАЗВОЈНОМ СКУПУ (120 РЕЧЕНИЦА).	87
СЛИКА 5-4 АНОТИРАНИ СЕГМЕНТИ ТЕКСТА ПРЕМА БРОЈУ ПОНУЂЕНИХ МОГУЋИХ ЗНАЧЕЊА У ИЗДВОЈЕНОМ ТЕСТ СКУПУ (300 РЕЧЕНИЦА).	88
СЛИКА 5-5 РАДНО ОКРУЖЕЊЕ INSERTION ЗА АНОТАЦИЈУ КОРПУСА.....	90
СЛИКА 5-6 БРОЈ ТАЧНО РАЗРЕШЕНИХ СЛУЧАЈЕВА НА РАЗВОЈНО-АНАЛИТИЧКОМ СКУПУ ОД 120 РЕЧЕНИЦА	94
СЛИКА 5-7 БРОЈ ПОГРЕШНО РАЗРЕШЕНИХ СЛУЧАЈЕВА НА РАЗВОЈНО-АНАЛИТИЧКОМ СКУПУ ОД 120 РЕЧЕНИЦА.	95
СЛИКА 5-8 БРОЈ НЕДОДЕЉЕНИХ СЛУЧАЈЕВА НА РАЗВОЈНО-АНАЛИТИЧКОМ СКУПУ ОД 120 РЕЧЕНИЦА.	96
СЛИКА 5-9 БРОЈ ТАЧНО РАЗРЕШЕНИХ СЛУЧАЈЕВА НА ИЗДВОЈЕНОМ ТЕСТ СКУПУ ОД 300 РЕЧЕНИЦА.	98
СЛИКА 5-10. БРОЈ ПОГРЕШНО РАЗРЕШЕНИХ СЛУЧАЈЕВА НА ИЗДВОЈЕНОМ ТЕСТ СКУПУ ОД 300 РЕЧЕНИЦА.	99
СЛИКА 5-11 БРОЈ НЕДОДЕЉЕНИХ СЛУЧАЈЕВА НА ИЗДВОЈЕНОМ ТЕСТ СКУПУ ОД 300 РЕЧЕНИЦА... ..	100
СЛИКА 5-12 ПРИМЕР ПОРАВНАЊА ЗНАЧЕЊА ИЗ ОПИСНОГ РЕЧНИКА И СИНСЕТОВА СРПСКОГ ВОРДНЕТА ЗА ЛЕМУ „АДМИНИСТРАЦИЈА”.....	113
СЛИКА 5-13 РАСПОДЕЛА МОДЕЛСКИХ ОЗНАКА У ПУНОМ СКУПУ	115
СЛИКА 5-14 РУЧНА ОЦЕНА МОДЕЛСКИХ ОЗНАКА У ПРОВЕРЕНОМ ПАКЕТУ	116
СЛИКА 5-15 КЛАСЕ КОМЕНТАРА АНОТАТОРА	117
СЛИКА 5-16 КОРИСНОСТ МОДЕЛСКОГ ОБРАЗЛОЖЕЊА ПРЕМА ОЦЕНИ АНОТАТОРА.....	118
СЛИКА 5-17 ПРИМЕР ИЗРАЧУНАВАЊА СЕНТИМЕНТА РЕЧЕНИЦЕ У S7-WSD КОНФИГУРАЦИЈИ.....	119
СЛИКА 5-18 РАДНИ ПРИМЕР ПОРЕЂЕЊА МАШИНСКОГ ПРЕВОДА БЕЗ ЗНАЧЕЊСКЕ НАПОМЕНЕ И СА ЗНАЧЕЊСКОМ НАПОМЕНОМ.....	122