

АЛГОРИТАМСКА ИНТЕРТЕКСТУАЛНОСТ ЗА КЛАСИФИКАЦИЈА НА ВАРИЈАНТИ НА НАРОДНИ ПЕСНИ И ПРОНАОЃАЊЕ СПОДЕЛЕНИ ТЕКСТУАЛНИ ФОРМУЛИ

Огнен Спироски

Природно-математички факултет, УКИМ – Скопје

Датум на прием: 17.5.2025

Апстракт: Во овој труд е претставена успешна примена на алгоритми за наоѓање на ист програмски код (текст) во колекција изворни кодови, за препознавање на делници со ист текст во македонски народни песни, нивно пронаоѓање и кластерирање по сличност и класификација на текстуални варијанти на песни. Ова е прв пример во литературата за алгоритамска интертекстуалност во народна поезија, фолклористика и етномузикологија, на истражување на варијанти на песни и текстуални формули, со методи развиени за детекција на плагијаризам во софтвер.

Во трудот има преглед на литературата и на истражувањата, поврзани со интертекстуалност и варијанти кај народни песни, кои вродиле нови класификаторски модели и теории во науката, а и голем број публикации што содржат варијанти на песни. Алгоритамската интертекстуалност овозможува нови дигитални анализи на големи корпуси песни, варијанти и формули.

Се анализираат два корпуси песни со софтверската алатка JPlag v5.1. Првиот корпус содржи 51 текстуална варијанта од 6 фамилии песни, кои се веќе изучувани, а алгоритмите ги кластерираат соодветно на фамилиите. Вториот корпус е голема, но некурирана колекција на 1.467 текстови песни од веб-страницата www.pesna.org. Алгоритмите во него пронаоѓаат и групираат варијанти на песни, а и многу споделени текстуални формули, кои нудат и интересни увиди во интертекстуалноста и во формулаичната природа на народната поезија.

Клучни зборови: интертекстуалност, формули, варијанти на песни, класификација, фамилија песни, детекција на плагијаризам во програмски код

Вовед

Поезијата и музиката се двата уметнички медиуми во кои народната песна живее и доживува еволуција, што вродува и со текстуални и со музички варијанти. Варијантите на народната песна можат да се анализираат и со примена на генеалошки парадигми, пристап што го појаснува споделеното (генетско) наследство во еволуцијата на варијантите. Терминот *фамилија песни* е дефиниран за блиските мелодиски варијанти, меѓутоа принципиелно важи за кои било информации (генетски) споделени меѓу песните, вклучувајќи го текстот на песната (Bayard 1950; Спироски et al. 2024).

Со *интертекстуалност* е означен феноменот на поврзаност на два (и повеќе) текста преку нивна текстуална и симболична сличност, како при цитирање или при употреба на фрагменти од друго дело, или

при какво било референцирање на екстерни симболи како извор на значење. Теоријата на интертекстуалност, дефинирана од Кристева и Барт околу 70-тите и 80-тите години од 20 век, развила теоретски и практичен интерес во референцирање и во толкување текст, базиран на структурализмот и на семиотиката (Ќулавкова 2021; Chandler 2022). Интертекстуалноста е дел и од фолклористиката, и како теорија и како феномен при истражувањата, како тие на народните песни (Димовска 2014, 2022; Бурић 2016).

Информатичките и податочните науки овозможуваат разни квантитативни и статистички анализи на текст, што доведе до нивна масовна примена и е секојдневие за пребарување на интернет. Пронаоѓање и работа со интертекст во дигитални архиви е една примена за која постојат разни алгоритамски методи, од кои некои во доменот на интертекстуалност, формулаичност и компаративна поезија (Suchecka & Gasiglia 2021). Во полињата на Digital Humanities и на Computational Literary Studies, интертекстуалноста најчесто е од нејзиниот наједноставен теоретски облик, пронаоѓање на ист или сличен текст (Text Reuse Detection), и е со фокус на практична методолошка примена на информатичка обработка на податоци, пред сè, за споредување на текст. Обично се применува врз големи дигитализирани корпуси од стари и историски текстови (Kallio et al. 2022; Düring et al. 2023), а почнува и употреба на семантички збогатени корпуси и модели од вештачка интелигенција за препознавање интертекст според: значење, метрика, стилometriја и др. (Schöch et al. 2023; Sarv & Järv 2023). Во последниве години, алгоритамската интертекстуалност сè повеќе се применува врз податоци од фолклористиката и од поезијата, како на пример за наоѓање интертекстуалност и формулаичност со биоинформатички секвенциски алгоритми за израмнување текст, во 3 корпуси со финска и естонска орална поезија, со вкупно 283.571 народни песни (Janicki 2023; Janicki et al. 2024).

Македонските народни песни се забележани со варијанти уште во најраните публикации од 19 век. Иако текстуалните варијанти, наспроти музичките варијанти се повеќе запишувани, објавувани и истражувани, сепак постојат малку компаративни анализи на варијантите на текстовите и на нивните систематизации. Варијантите најчесто биле анализирани рачно и квалитативно: фолклористички, историски и сл., поретко содржински: лингвистички, музички и сл. (Спироски 2025).

*

Во овој труд се претставува успешна употреба на алгоритамски методи за наоѓање на ист текст во изворни кодови на програми (Prechelt et al. 2002; Sağlam et al. 2024), за пронаоѓање интертекстуалност во македонски народни песни и нивно групирање

(кластерирање) според текстуална сличност, т.е. според класификација на варијантите на песните. Резултатите нудат увид и во тоа кои се текстуалните формули во еволутивно конзервираните, односно споделените делници со идентичен текст, низ песните и варијантите. Ова претставува прва примена во досега објавената научна литература, на методи развиени за детекција на плагијаризам во програмски код, врз податоци од: народна поезија, фолклористика и етномузикологија.

Во трудот прво се валидираат методите врз варијанти од 6 (шест) народни песни за кои веќе постојат објавени истражувања за нивни сличности, односно за нивната припадност во една „фамилија песни“ според мелодијата и/или според текстот. Алгоритмите за детекција на плагијаризам успешно пронаоѓаат текстуални сличности и формули, т.е. споделени фрагменти текст, и во парови песни, и во групи од песни (Сл. 1–3, 8).

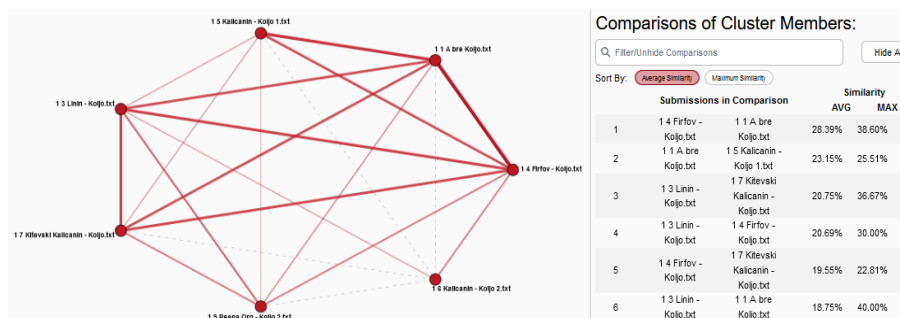
1	Дафина болна легнала,	1	Дафина болна легнала
2	над глава јој три доктора,	2	над глава до три доктора.
3	Болна е, болна преболна,	3	Болна е, болна ќе умре,
4	ќе умре млада Дафина.	4	ќе умре млада Дафина.
5		5	
6	Ој мајко, мила мајчице,	6	- Мајко ле, стара мајчице
7	викни го моето либе,	7	повикај ми го либето,
8	два збора да му прозборам,	8	два збора да му прозборам,
9	до срце да го изгорам.	9	до срце да го изгорам.
10		10	
11	Дур не помине година,	11	Дури не мине година,
12	да носи црна кошула,	12	бела руба да не носи.
13	дур не поминат две години,	13	бела руба да не носи,
14	друго моме да не љуби.	14	друго либе да не љуби.

Сл. 1: Споредба на 2 варијанти во JPlag (Каракаш 1995; www.pesna.org), истите делници (интертекстот) се означени во боја; мин. 3 исти збора ($t=3$)

Алгоритмите за кластерирање успешно групираат варијанти на текстови на песни во заеднички кластери (Сл. 2), соодветни на фамилии песни, но даваат различни множества кластери со варирање на параметрите. Резултатите се точни и информативни, но нелинеарни при варирање на параметрите, па бараат додатно (рачно) истражување и бирање на значајни резултати (кластери со варијанти) од различните множества кластери добиени при варирањето. Ова е очекувано во истражувања со вакви алгоритми, а слични рачни и автоматизирани процеси на рафинирање резултати се опишани во литературата.

По успешната валидација на алгоритмите врз курирани текстови од однапред познати фамилии песни, претставуваме успешна примена врз голем корпус некурирани текстови народни песни, достапен на интернет, со 1.467 текстови, најдени на веб-страницата www.pesna.org. Резултатите вклучуваат пронаоѓање сличности во текст од песни, групирање песни според сличност и наоѓање потенцијални текстуални варијанти и фамилии песни. Репортирано е за резултатите од

варирањето на параметрите на алгоритмите и нивниот ефект врз бројот на идентификувани слични текстови и врз групациите текстови во кластери. Резултатите се илустрирани и со слики и со дискусија врз мал број примери од наодите, какви што се некои познати песни и нивните варијанти, а и со некои текстуални формули.



Сл. 2. Кластер со 7 варијанти на „А, бре, Кољо“ (резултати во JPlag; лево: визуелни сличности како врски; десно: нумерички сличности)

Интертекстуалност, бриколаж и варијантност во фолклорот

Варијантите на народните песни содржат значителна до речиси целосна интертекстуалност, појава дискутирана уште од најраните собирачи, а истражувана низ разни теоретски и практични парадигми. Интертекстуалноста во фолклорот и во народната песна помогнале и за развивање нови модели и теории во науката. Интертекстуалноста е жив феномен, кој е и свесно применуван практичен пристап, и во препејувањето и во создавањето нови варијанти на постоечките народни песни, како и на нови песни (Верковиќ 1860; Поленаковиќ 1952; Хадиманов 1960; Линин 1976; Узунски 1991; Спироски и Спироски 2009; Ђурић 2013; Мојсова-Чепишевска 2013; Бурић 2016; Кузман 2021).

Практичниот интертекстуален пристап соодветствува со концепцијата на Леви-Строс за *бриколаж* и *бриколер*, соодветен за народниот уметник, кој реупотребува постоечки нешта што се нашле околу него (симболи, митови), и преку нив, со нивно колажирање и креативна реорганизација, создава варијантно дело, „нов“ мит (Lévi-Strauss 1974[1962]). Семиотските анализи на бриколажот покажуваат дека оперира преку неколку клучни трансформации: *додавање*, *бришење*, *заменување* и *транспозиција* (Chandler 2022). Тие се налик на трансформации и мутации на генетскиот код, со што го приближуваат бриколажот до еволутивните парадигми и методи, како кај: генетиката, меметиката, генеалогичката, филогенетиката и биоинформатиката. Преку тоа, дигитални истражувања во еволуцијата на интертекстуален бриколаж, т.е. на варијантите песни, се овозможени со методи од информатика.

Може да се каже дека симболите и текстот во народните песни се дел од културолошки генетски, односно меметски кодови (Dawkins 1976), чии (генетски) адели, односно мимови, се пренесуваат преку интертекстуален процес на бриколаж и се трансформираат во варијанти и во нови песни. Алгоритамските методи нудат нов начин дигитално да се препознае и да се истражува развојот на интертекстуалноста и на варијантите на народната песна и да се мапира нивниот генетско/меметско/формулаичен состав, да им се трасираат нивните мутации, рекомбинации и селекција низ „популација“ песни, т.е. да се мапира еволуцијата на нивната содржина и да се има увид во диверзитетот на варијантите на песните.

Потребата за подредување и групирање песни и варијанти вродило разни класификациски шеми и систематизации за народни песни, како литературната, модалната, лексикографската (Keller 1984). Од нив се издвојува граматичката, доработена од лексикографската од Барток, за потребите на анализа на балканскиот музички фолклор од теренските истражувања и звучните снимки на М. Пари и А. Лорд во Југославија, од кои е објавен препис и етномузиколошко истражување на 75 српско-хрватски песни (Bartók 1951).

Орално-формулаичната теорија на Милман Пари (1971) и Алберт Лорд (2000 [1960]) ги истражува повторливите метрички јазични обрасци (*интертекстуални формули*) во песните и нивната улога во оралната композиција, пред сè, на долгата епска поезија. Пари и Лорд (со Барток) направиле фундаментални студии и за варијантноста, и според видот и обемот на фолклористичкиот материјал, и според израмнувањето на стиховите на варијантите за споредување на текстовите и на формулите од песните, но пред сè, во квантитативните анализи на текст и на формулаичност. Нивната теорија и пристап имаат доста успеси, но и критики (Ljubinković 1991), од што следат разни нивни проширувања и доработувања (Foley 1988). Ова е рефлексција на тренд на еволуција на фолклористичкиот поглед на варијантите, од потрага по „оригинален“ архетип, во анализа на флуидноста на фолклорот (Спироски 2025).

Интертекст, неинтертекст и текстуални формули во еволуција на варијанти на народни песни

Концептот на јазична формула, кај варијантите на песни, во основа подразбира интертекстуалност бидејќи варијантоста им е дефинирана со имање споделен текст, односно споделени формули и теми.

Споделениот текст и формулите меѓу варијантите на една песна ги претставуваат *еволутивно конзервирани* делници, најстабилниот и најпрепознатлив текст, со најмалку промени, односно мутации. Може да се каже дека *интертекстуалноста*, односно споделените

текстуални формули меѓу варијантите, се дефинирачкиот текстуален *џенейски скелет* на фамилијата песни (види: Сл. бр. 1, 3 и 8, обоен е само интертекстот).

Остатокот од текстот што не е споделен и не е сличен меѓу варијантите (*неинтертекст*), што е локален и специфичен за една или за мал број варијанти, го претставува *еволутивно дивергентниот, агајтивниот и варијантен* дел од песната, подложен на еволутивни притисоци како мутација, селекција итн. Овие локални, несподелени содржини на песната, со своите варијации, обично носат повеќе информации и поголем диверзитет на зборови и фрази отколку споделените формули. Содржините на овој неинтертекстуален дел, каде што се манифестира варијантноста, може да бидат и покорисни за истражувачите отколку интертекстот, па вреди да бидат разгледани и тие содржини (како „инверзен“ резултат од пронајдениот интертекст).

Алгоритамскиот пристап, употребен во овој труд, овозможува автоматизирано пребарување во голем корпус песни и наоѓање на парови слични текстови (и со минимална и со значителна количина интертекст), со веднаш видлив варијантен неинтертекст, за лесно означување, односно екстрахирање на различните зборови за понатамошна обработка, и анализи, како тие на Поленаковиќ (Спироски 2025).

За варијантите на песната „Распукала Шар Планина“

Во зборникот на Миладиновци во 1861 првпат е објавен текст на „Падна Шар планина“, која е веројатно првата македонска народна песна во 19 век објавувана и со повеќе варијанти во рамките на една публикација и со компаративна анализа на стиховите на варијантите. Беа пронајдени над 15 изданија на песната во 19 век, од кои дури 3 со повеќе варијанти во иста публикација, а барем 10 се објавени во периодот од 1889 до 1895 година. Некои се реобјавувања од стари публикации и обично се со референција. Има објавено и референции до варијанти во други публикации.

Некои од анализите на варијантите на песната за Шар Планина се, пред сè, дел од балканските пропаганди за Македонија и за Македонците, но се интересни и во контекст на разбирањата на варијантноста и на интертекстуалноста. За неодржаното аргументирање на Гопчевиќ (1889) дека објавената варијанта на песната, наведена во зборникот на Миладиновци, е „фалсификат“ и објавувајќи „вистинска“ за која вели дека е српска, Јуриновиќ (1892) има публикуван противаргумент, што успешно го демантира со лингвистичка анализа на песната и со повеќе објавени варијанти.

Варијанти на песната се објавувани и се истражувани и во 20 и во 21 век, со над 10 публикации со варијанти (Василевиќ 1953; Фирфов и Симоновски 1962; Спироски и Спироски 2009; Јаневска

2024). Во анализите на неколку објавени варијанти се дискутираат текстуални и мелодиски сличности, во контекст на анализата на еволуирање во песна од мешовит, селско-градски тип, од 19 до 21 век (Stojkova Serafimovska 2020; Тончева 2021).

Текстови на некои варијанти или аудиоснимки на песната можат да се најдат и онлајн, вклучувајќи и на lyrics Web-sites. На www.pesna.org за „Се навали Шар Планина“ има текст и 9 аудиоснимки, од кои дури 3 се верзии снимени од Александар Сариевски. Постојат многу снимки и преработки од разни музичари што не се на веб-страницата. Интересно, на веб-страницата постои и авторска песна („Распукала Шар Планина“ од Г. и Л. Кукиќ), што со тој стих интертекстуално ја референцира народната песна.

Во овој труд не се употребени сите најдени текстови на песната.

Податоци

Користиме 2 корпуси податоци. Првиот корпус е за валидација на алгоритамското препознавање сличности на варијанти на текстови од однапред познати фамилии песни и нивно групирање во кластери. Вториот е голем онлајн корпус, за експериментално барање сличности.

Корпус #1 – 6 фамилии песни

Корпусот #1 содржи вкупно 51 варијанта од 6 познати фамилии песни. Бидејќи некои се веќе истражени за сличности во други трудови и се познати, се соодветни за валидација на алгоритамската методологија за класификација во фамилии според кластерирање по сличност на текст.

1. А бре Кољо, мамин Кољо – 9 текстови
2. Дафина болна легнала – 6 текстови
3. Излегла Јана по поле – 10 текстови
4. Марика мома убава – 7 текстови
5. Море, се разболе млад Никола – 2 текста
6. Распукала Шар Планина – 17 текстови.

Првите 5 фамилии се алгоритамски анализирани и објавени како Додаток и сите мелограми со референци до нивните публикации (Спироски et al. 2024). Во овој труд ги користиме песните од истите референции, заедно со нови варијанти и од други извори. За шестата песна, за „Шар Планина“, во трудов е даден краток преглед на литература со варијанти и анализи на песната.

1	Излегла Јана по поле,	1	Излегла Јана по поле,
2	Да види Јана полето,	2	Да види Јана полето,
3	Да види Јана полето,	3	Да види Јана полето,
4	Дали е поле стасано,	4	Да ли е жито фтасало,
5	Да фанит Јана аргати,	5	Да ли е жито фтасало,
6	Аргати млади жетварки.	6	Да вати Јана аргати,
7		7	Да вати Јана аргати,
8	На сретшта иде турчинче	8	Аргати, млади жетварки.
9	Турчинче младо спаивче.	9	
10	Турчинче младо спаивче,	10	Сретна ми Јана туѓинче,
11	Спаивче на Јана говори:	11	Туѓинче, младо Турчинче
12	Дигни си Јано тулбенче,	12	Туѓинче, младо туѓинче,
13	Да ти го видам лицето.	13	На Јана вели, говори:
14		14	
15	Бегај бре младо спаивче,	15	Одвржи Јано, тулбенот,
16	Јас имам моје бекарче,	16	Да ти го видам черето,
17	Овчарче младо чобанче.	17	Да ти го видам черето,
18		18	Да ли прилича за мене?
19		19	
		20	Бегај, бре Турче, туѓинче,
		21	Ја имам моје бекарче,
		22	Ја имам младо бекарче,
		23	Овчарче, младо чобанче!

Сл. 3: „Излегла Јана по поле“ (Хаџи-Манев 1955; Васиљевић 1953) во JPlag

Корпус #2 – www.pesna.org

Многу варијанти на текстот и на мелодијата на македонски песни се зачувани во звучни записи. Во евиденција на варијантите во снимките на онлајн архивата www.pesna.org има барем 1.457 текстови на песни, кои имаат 3.988 звучни записи. Ова е некурирана колекција, има и новокомпонирани народни песни, па е пример за „real world scenario“ тип на податоци. Соодветна е за експериментална анализа со применетата методологија за алгоритамското барање на текстуални варијанти и формули.

Песните обично се претставени со само еден текст, најчесто препишан од еден од повеќе звучни записи (изведби) за истата песна. Во корпусот има и мал број варијанти на текст, независно препишани од доволно различни варијанти. Околу 40 текстови на песните содржат уште една варијанта, запишана последователно по првиот текст. Некои текстови содржат додадени забелешки, како дека постои слична песна (варијанта) на веб-страницата, или за потеклото на песната и сл.

Текстовите од веб-страницата беа најчесто искористени целосно како што се објавени. Беа отстранети мал број додадени забелешки во текстот, што во некои случаи резултираше со кластерирање текстови според забелешката. Текстовите би можеле уште да се процесираат, првенствено да се одвојат како посебни, текстуалните варијанти објавени под една песна. Меѓутоа, ова и други процесирања на содржината на текстовите се оставени за наредни истражувања, пред

сè, за да се презентира анализа на содржината на веб-страницата, како што била најдена, со минимум уредување.

Методи

Секој метод за барање сличности на текст носи свои предности и недостатоци. Некои методи подобро се справуваат со наоѓање сличност и покрај (мали) разлики во текст, како биоинформатичките методи за израмнување секвенции (Janicki 2023; Janicki et al. 2024). Други методи нудат голема брзина и се корисни за брзо профилирање големи колекции текст, но не даваат увиди во тоа какви се сличностите на текстовите (Schöch et al. 2023).

Алгоритамскиот пристап за детекција на ист текст, користен во овој труд (Prechelt et al. 2002; Sağlam et al. 2024), автоматски ги пронаоѓа и ги означува споделените текстуални формули, т.е. еволутивно конзервираниот интертекст на парови песни/варијанти (Сл. 1, 3, 8). Врз база на квантитативна сличност на најдениот интертекст, автоматски се пресметува и кластерирање на текстовите/варијантите (Сл. 2, 7). Еволутивно дивергентниот („инверзен“ неинтертекст) не е означен, но неговата должина влијае на двете пресметани сличности. Иако во компаративниот приказ на парот песни, неинтертекстот е видлив во контраст со интертекстот, сепак, со применетата софтверска алатка JPlag, мора „рочно“ да се прегледува, да се бира, да се зачувува и додатно да се анализира. Затоа во овој труд ќе се задржиме само на општи коментари и увиди за неговата содржина.

Применетите алгоритми се дел од JPlag v.5.1.0, софтвер со отворен код за детекција на плагијаризам во програмските кодови. JPlag¹ е дизајниран да работи со изворен код од многу програмски јазици, а и со обичен текст, како што е применет во овој труд.

JPlag нуди две главни методолошки можности:

- споредување на 2 текстуални датотеки (мерење сличности);
- препознавање на групации од најслични датотеки меѓу себе (кластерирање по текстуална сличност).

И двата методи имаат свои параметри што влијаат врз работата и врз резултатите, т.е. колку и какви сличности и кластери се препознаени.

JPlag нуди репортажа, видлива во веб-пребарувачот со веб-апликацијата (Web app), со листа на сличности (Сл. 7), преглед на резултати од споредбите (Сл. 1 – 3, 8) и статистичка дистрибуција на процентуалната сличност (Сл. 4 – 6). Ова е информативно за сличностите меѓу текстовите и барањето варијанти.

¹ JPlag може да се најде онлајн на: <https://github.com/jplag/JPlag> [Пристапено на 14.8.2024].

Сличност на пар текстови (Споредување 2 датотеки во JPlag)

JPlag има 2 мерки за сличност меѓу пар од 2 текста/датотеки:

- просечна сличност (паровите се споредуваат во цела должина);
- максимална сличност (се споредуваат само истите делници, подобра за текстови со разлики во должини и за фрагменти).

Сличноста се пресметува користејќи *џокенизација*, одделување само на зборовите од текстот како независни, помали, основни единици за споредба – *џокени*, што овозможува препознавање на споделени секвенции на исти токени, односно исти делници текст во 2 датотеки. Означениот текст во боја во репортажата се истите токени (Сл. 1, 3, 8).

Во барање значајни сличности меѓу 2 датотеки главен параметар е *минимум број џокени* *t*, со кој се регистрираат само делници што во двата текста споделуваат минимум *t*-последователни исти зборови. Споделените делници можат да содржат повеќе токени од минимумот. Ефектите на овој параметар (*t*) се истражувани во резултатите подолу.

Благодареејќи на ова, JPlag добро се справува со некои аспекти на споредување текст, како специјални карактери (запирки, точки ...), поделби низ редови, расцепкани фрагменти текст (стихови во различен редослед). Поради текстуалното точно и егзактно препознавање на иста секвенција токени, JPlag не препознава слични и речиси исти зборови. Ова значи дека „се пропуштени“ многу потенцијални сличности, некогаш поради разлика од само една буква во збор. Од друга страна, сите најдени сличности се сигурни резултати, со сосема исти зборови.

Кластерирање на слични текстови во JPlag

JPlag има можност за кластерирање со 2 алгоритми, и за репортирање листа на кластери – групации од најслични текстови. Има 2 визуелни прикази на кластерите, мрежа/граф (Сл. 2) и радар. Кластерите имаат и референции до помалку слични текстови, што не се дел од кластерот. Ова овозможува класификација на варијанти според кластери, како и барање текстуални формули и варијанти во големи корпуси текст.

Двата алгоритми за кластерирање во JPlag се:

- спектрално кластерирање;
- агломеративно кластерирање.

Алгоритмите оперираат врз истите квантитативни сличности, а имаат свои специфични метрики и параметри, а дел од нив се истражувани во резултатите подолу.

Резултати

Истражувани се резултатите од варирање на параметрите и на алгоритмите врз бројот на најдени сличности и групации во кластери. Посебно се истражувани двата корпуси со различни цели и варирања.

Највлијателен параметар врз пресметките на сличност е бројот на минимум (последователни) токени, t . Го истражуваме варирањето на параметарот t врз споредбите и во двата корпуси, во рангот $t = \{2..6\}$

- Со мал мин. број токени t се регистрираат повеќе сличности. Сличностите некогаш не се толку значајни (= помалку ист текст).
- Со зголемување на мин. бр. токени t има помалку сличности. Сличностите обично се многу посигурни (= повеќе ист текст).

Корпус #1 – 6 фамилии песни

Првиот корпус е истражуван на два начини, според избор на текстови:

- само текстови од поединечната фамилија песни (табели 1 – 2). Нема кластерирање само за фамилијата „Никола“ бидејќи со 2 текста можна е само 1 споредба, што е дадена процентуално.
- сите текстови од сите 6 фамилии песни заедно (табели 3 – 4).

И на двата начини е вариран параметарот t , како и двата алгоритми за кластерирање (спектрален, агломеративен). Во табелите 1 – 4 се дадени бројките на текстуални варијанти што се групирани во кластер за соодветната фамилија, како и вкупниот број текстови за фамилијата. Бројот на кластерирани од вкупниот број варијанти, би можело да се третира како процент на „успешноста“ на алгоритмите да класифицираат текстови на познати фамилии песни. Најголема „успешност“ има за ниските вредности $t = \{2, 3\}$ (види: табели 1 – 4).

Ниеден алгоритам не успева целосно да ги покрие сите примери фамилии за да биде употребен автоматизирано за класифицирање текстови во фамилии песни, но сепак, алгоритмите успешно класифицираат текстови во фамилии и откриваат битни сличности. Резултатите се информативни и можат да се употребат во истражувања.

Песна	Бр. вариј.	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
Никола	2	60.47%	51.16%	46.51%	34.11%	26.36%
Дафина	6		4		4	
Марика	7					
Кољо	9	7	7	7	6	4
Јана	10	6	10	6	6	
Шар планина	17	16	11	10	9	11

Табела 1: Spectral clustering – Поединечни фамилии песни

Песна	Бр. вариј.	t = 2	t = 3	t = 4	t = 5	t = 6
Никола	2	60.47%	51.16%	46.51%	34.11%	26.36%
Дафина	6					3
Марика	7		7	5	7	7
Кољо	9	3				
Јана	10				5	3
Шар планина	17	12	10	9	9	6

Табела 2: Agglomerative clustering – Поединечни фамилии песни

Песна	Бр. вариј.	t = 2	t = 3	t = 4	t = 5	t = 6
Никола	2		3			
Дафина	6			3	3	3
Марика	7	7	3	3	3	
Кољо	9	4	3	3	3	
Јана	10	5+8	3	3		
Шар планина	17	3+3+3+4	3+3	3	3+3	3

Табела 3: Spectral clustering – Групирани сите фамилии – Збировите претставуваат повеќе најдени кластери, со бројот на песните во нив

Песна	Бр. вариј.	t = 2	t = 3	t = 4	t = 5	t = 6
Никола	2					
Дафина	6	6	6	5	5	3
Марика	7	7	7	7	7	3+4
Кољо	9	3				
Јана	10	9	7	5+3	5	3
Шар планина	17	14	12	9	9	6

Табела 4: Agglomerative clustering – Групирани сите фамилии

Корпус #2 – www.pesna.org

Се бараат постоечки сличности меѓу текстовите без априори знаење на тоа какви сличности и варијанти ќе се најдат во корпусот. Успешно се наоѓаат многу сличности во голем број текстови, но има потреба од филтрирање на многу од сличностите како нерелевантни. Бројот на максимум споредби е еден филтер во JPlag, што одредува граница колку од првите најслични споредби ќе бидат во репортажата.

Истражувани се 2 алгоритми, 4 метрики за 1 алгоритам, минимум број токени $t = \{2, 3, 4\}$, максимум споредби = $\{500, 1000, 2000\}$.

Резултатите на сликите 4 – 6 се за 3 вредности на t , а ги покажуваат дистрибуциите на процентуални сличности кај паровите споредби. Има мал број споредби (14, 9, 7) со над 50% сличност, што се многу веројатни варијанти.

Од најсличните текстови, голем дел не се во кластери, за сите вредности на t . Бројот на некластерирани слични парови варира – од 20% до 60% од текстовите од топ најслични не се во кластери, а се многу веројатни варијанти.

Во топ 30 најслични текстови, речиси сите се варијанти на песни. Само 2 текста се слични според рефрен („оф, аман, аман“, „тра ла ла“).

Максимум споредби	Clustering / metric	$t = 2$	$t = 3$	$t = 4$
500	Default (Spectral)	117	166	126
1000	Default (Spectral)	285	365	250
2000	Default (Spectral)	287	530	297
1000	Max (Spectral)	238	310	340
1000	Min (Spectral)	223	344	253
1000	Average (Spectral)	228	400	280
1000	Intersection (Spectral)	227	224	259
1000	Agglomerative (def.)	86	43	21
2000	Agglomerative (def.)	87	43	21

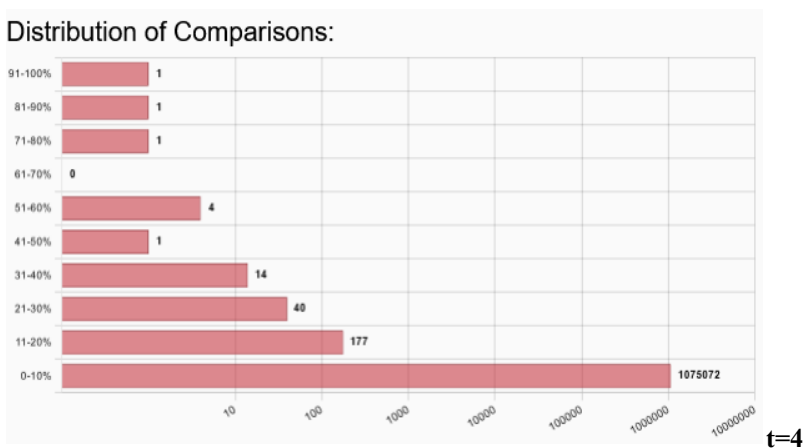
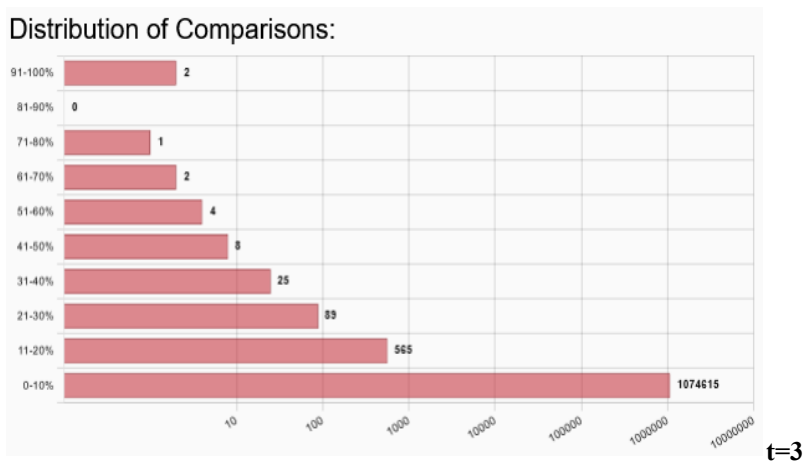
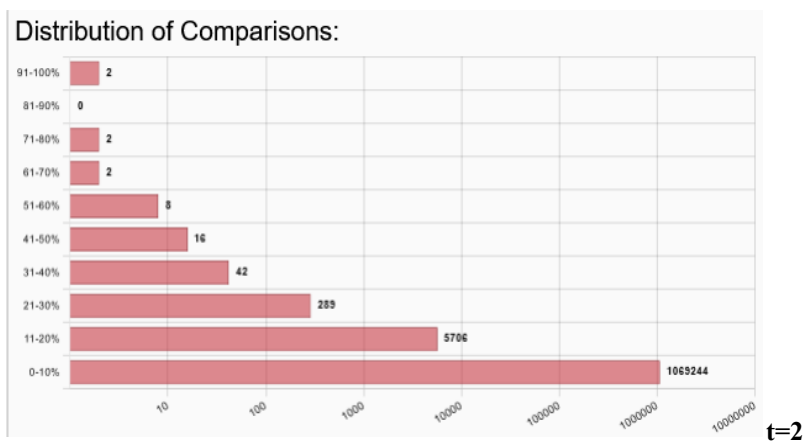
Табела 5: Вкупен број парови споредби во сите најдени кластери

Во табелата 5 се претставени варирања на параметрите за двете кластерирања со default-метриката, а и за 4 метрики на спектралното кластерирање. Има различни множества на кластери при варирање и токенизација. Има значително повторување на исти кластери низ вредностите на t , но има и уникатни кластери за секоја вредност на t .

Спектралното кластерирање со сите метрики наоѓа повеќе сличности отколку агломеративното, кое дава помалку, но посигурни сличности и е под слабо влијание од бројот на максимум споредби.

Просечна сличност во рамките на еден кластер е максимум околу 30%.

Многу од поголемите кластери „се разводнети“, т.е. со слаби или незначайни сличности, што го намалуваат просекот на кластерот. Овој феномен се зголемува со растење на бројот на максимум споредби.



Сл. 4 – 6: Дистрибуции на процентуална сличност на споредби на текстови од www.pesna.org, за 3 токенизации ($t = 2 - 4$), max-споредби = 1000

Дискусија

Токенизацијата има најмногу влијание на резултатите, пред сè, на поединечните споредби, т.е. сличности, но и во кластерирањето. Идеално е да се истражат повеќе резултати преку варирање на параметрите бидејќи постојат уникатни разлики во секој резултат. Главно, добри резултати се добиваат околу мин. број на токени од $t = 3$.

Спектралното кластерирање е типично посеопфатно и малку поинформативно од агломеративното, и е посоодветно за истражување на поголеми и на непознати содржини. Агломеративното дејствува подобро за групирање кога има повеќе варијанти во диверзни колекции. Идеално е да се употребат и двете кластерирања – за повеќе увиди во специфичностите што ги имаат во резултатите и во нивната синтеза.

Кластерирање на варијанти на песни

Алгоритмите успешно групираат текстуални варијанти соодветни на нивната фамилија песни. Иако ретко се групираат повеќето варијанти од фамилијата во еден кластер, тоа е соодветна рефлексивност на разликите во содржините на употребените текстови.

Се добиваат нелинеарни резултати при варирање на токенизацијата – снемивање или намалување, па растење на бројот на песни во еден кластер.

За една фамилија песни, сосема фалат кластери кај поединечните фамилии песни со спектралното кластерирање. Иронично, тоа е за песните со најслични текстови, од фамилијата „Марика“. Во групирани фамилии, спектралното кластерирање наоѓа кластери и од „Марика“. Агломеративното, во групирани фамилии, е најуспешно за овој пример.

Ова укажува на „потреба“ од т.н. надворешна група (доволно различен примерок), бидејќи некогаш алгоритмите кластерираат „посигурно“ со колекција од диверзни текстови.

Во групирани фамилии песни (Табела 4), агломеративното обично групира повеќе варијанти од спектралното кластерирање. Но, спектралното нуди други увиди, преку наоѓање на повеќе различни кластери за истата фамилија (Табела 3), поконзистентно е за варирање на t , а и групира повеќе песни при поединечните фамилии (Табела 1).

Класификација во фамилии песни преку кластерирање

Алгоритмите за кластерирање можат да се употребат во студии за класификација на текстови на песни во фамилии. Откриваат слични

групации, еволутивно конзервиран интертекст и текстуални формули. Индиректно, откриваат и варијантен, дивергиран неинтертекст. Меѓутоа, не можат да се употребат за добивање на една „целосна“ слика на сличности, со само една вредност на параметрите, која не е доволна за целосен увид во ширината на резултатите. Потребно е варирање на параметрите за истражување на нелинеарните резултати.

Тука се поставува едно фундаментално прашање:

Што ѝрејсѝавува усѝешна класификација во фамилија ѝесни?

Разгледуваме три случаи:

1. Еден кластер – целосна групација на (речиси) сите варијанти.

Открива пошироки врски, квантитативно послаба сличност (намален просек), соодветен за класификација на варијанти.

2. Повеќе кластери – со различни (помали) комбинации од варијантите од истата фамилија.

Открива уникатности во групирањата, квантитативно појака сличност по кластер, полесни за толкување и анализа.

3. Некластерирани споредби – со голем процент на сличност, уникатни парови текстови што немаат други слични на нив.

И поединечни парови текстови што не се дел од кластер, можат да се третираат за фамилија од 2 варијанти. Останува на истражувачот (рочно) да ги избере значајните парови.

Алгоритамските истражувања на класификација на варијанти треба да ги имаат предвид сите овие можности, а за формирање (научни) увиди, резултатите од информатичките анализи треба да бидат (рочно) прегледани од експерти од соодветна (научна) област, како од: фолклористика, лингвистика, етномузикологија и др.

Ќе го споредиме спектралното кластерирање за поединечна фамилија песни, на 2 (од 4) кластери со по 3 текста за „Шар Планина“ (Сл. 7), со агломеративно, со еден голем кластер од 14 текстови (Таб. 4). Двата мали кластери имаат висока просечна сличност, од 83,17% и 79,81%, додека големиот кластер има мала просечна сличност, 42,35%. Вакви помали, посигурни групации се и полесни за толкување.

Разгледувајќи ги малите кластери е информативно и за врските меѓу варијантите и за нивната сличност, еволуција и генеалогичка. Во првиот кластер (Сл. 7), текстот од Сариески е веројатно појдовна верзија за тие на www.pesna.org и на WikiMedia, со кои има најголема сличност, и кои веројатно се преписи токму од снимката на Сариески. Во вториот кластер, текстовите на Бадев и Димчевски се речиси идентични, а веројатно нивна појдовна верзија е таа од Хациманов, со која и двете имаат блиски сличности, а е објавена пред нив.

	Submissions in Comparison		Similarity		Cluster
			AVG	MAX	
1	6 11 PesnaOrg - Se navali Shar.txt	6 13 - Lyrics Sarievski - Se navali Shar.txt	91.17%	92.81%	3  83.17%
2	6 12 Wiki - Raspukala Shar.txt	6 11 PesnaOrg - Se navali Shar.txt	83.52%	85.07%	3  83.17%
3	6 12 Wiki - Raspukala Shar.txt	6 13 - Lyrics Sarievski - Se navali Shar.txt	74.82%	77.61%	3  83.17%
4	6 15 - Dimcevski - Raspukala Shar.txt	6 14 - Badev - Raspukala Shar.txt	96.00%	97.78%	3  79.81%
5	6 5 Hadzi - Rasturi se Shar.txt	6 14 - Badev - Raspukala Shar.txt	72.18%	76.19%	3  79.81%
6	6 5 Hadzi - Rasturi se Shar.txt	6 15 - Dimcevski - Raspukala Shar.txt	71.26%	73.81%	3  79.81%

Сл. 7: Два (2) кластери на „Шар планина“ со 3 текста и висока сличност

Истражување сличности на непознат корпус – www.pesna.org

Експерименталната анализа на некурираниот корпус песни со JPlag покажа успешна примена на алгоритмска интертекстуалност, за барање: сличности, варијанти и нивна класификација (Сл. 8).

Најдени се многу споделени кратки текстуални формули низ разни песни, вклучувајќи рефрени и други мотиви. Такви секако има и во варијантите, но има многукратно повеќе во паровите споредби со слаба процентуална сличност. На пр. на Сл. 5, за $t = 3$, има 565 парови сличности од 11% до 20%, а само 14 парови со сличност од 51% до 100%. Секако има уште повеќе текстуални формули во сличностите под 10%, па затоа формулите се оставени за посебни истражувања и анализи.

Кратко резиме на резултатите:

- сличностите над 30% најчесто се варијанти на песни.
- постојат сличности и под 10%, што сепак се од варијанти.
- постојат слични варијанти што не се во кластер со фамилијата.
- се потврдени над 50 слични парови како варијанти на песни.
- се потврдени над 50 кластери што одговараат на фамилии песни. Бројот не е конечен. Потребно е рачно прегледување на многу резултати.
- корпусот има барем 100 варијанти, односно барем 7% од корпусот.

Заклучок

Во овој труд е претставено успешно препознавање текстуални сличности, формули и варијанти кај македонски народни песни, валидирајќи ја примената на JPlag и на неговите методи, во истражувања на алгоритмска интертекстуалност во народна поезија. Резултатите се информативни и корисни за познати и за непознати колекции песни.

1061.txt/1061.txt	55%	148.txt/148.txt	48%
1 Велешките ергени		1 Прилепските ергени	
2 Велешките ергени		2 Прилепските ергени	
3 каде одат сал пејат,		3 каде одат и пејат,	
4 по момите гледаат,		4 по момите гледаат,	
5 штом ги видат, ќе им свирнат		5 да ги видат, ќе свирнат.	
6 и со око ќе намигнат,		6 И со око ќе намигнат:	
7 леле, леле што се убави.		7 - „Леле, леле што се убави“.	
8		8	
9 А момите играат,		9 Стар бел дедо прислушал,	
10 гајтан веѓи дигаат,		10 бела коса исчешлал	
11 шага се подбиваат,		11 и на баба ѝ вели:	
12 ситно, ситно оро вијат,		12 - „Дај мори бабо да те бакнам јас,	
13 од ергени очи кријат,		13 и за образ да те штипнам,	
14 леле, леле што се убави.		14 леле, леле што си убави“.	
15		15	
16 Стари дедо ми седнал,		16 Кутра баба не гледа	
17 од далеку ми гледа		17 и на дедо му вели:	
18 и на баба ѝ вели:		18 - „Де бре старо, будало,	
19 - Елај бабо да те бацам,		19 пушти памет ти излегол,	
20 ич со рака не те факам,		20 пушти ѓавол ти навлегол,	
21 леле, леле што си убави!		21 леле, леле старо, будало“.	
22		22	
23 Стара баба сал се мршти		23 - „Пушти ѓавол ми излегол,	
24 и на дедо му вели:		24 пушти мерак ми навлегол,	
25 - А бре старо будало,		25 бабо мори за тебе,	
26 пушти ѓавол ти навлегол,		26 дај мори бабо да те бакнам јас,	
27 сиот памет ти излегол,		27 и со рака да те штипнам,	
28 леле, леле што си будала.		28 леле, леле што си убави,	
29		29 како солца да те макнам,	
		30 леле, леле што си убави“.	

Сл. 8: Две варијанти на песни најдени со JPlag на www.pesna.org

Валидирана е примената на кластерирање на текстови за класификација на варијанти во фамилии песни. Групирањата во кластерите се точни, но кластерите не се појавуваат секогаш конзистентно низ параметрите – има нелинеарни резултати и е потребно да се истражат повеќе варијации на параметри за формирање увиди.

Резултатите нудат можност за компаративна анализа на еволуција во варијантите, овозможувајќи брз преглед на еволутивно конзервиран интертекст, а и на диверзитетот на „мутациите“ на варијантниот текст.

Фолклорот на еден народ е налик на таписерија од споделени фрагменти на колективната меморија, а народната песна нуди брз пристап до ова споделено културолошко наследство. Алгоритамското препознавање на споделени текстуални формули низ народни песни овозможува правење дигитални мапи на варијанти и на споделени лиричко/меметски мотиви, како и идентификување на интертекст и формули што се еволутивно (нај)конзервирани, показател дека силно резонираат со народниот креативен нагон и со колективната свест.

Идна работа

Со JPlag v5.1.0 моментално не може да се направи кластерирање на сите примери, како што обично се прави при филогенетска анализа, ниту пак е соодветен за такви визуализации. Една можност би било да се експортираат резултатите на сличностите од JPlag и да се направи обид за нивно кластерирање со други софтвери и алгоритми.

Со дигитализирање на (целосен) корпус на сите досега објавени македонски народни песни, би се овозможило автоматизирано правење на дигитални мапи и индекси – на песни и на варијанти, подредени по сличност и по фамилии, на споделени текстуални формули и на нивната фреквенција и историја на употреба, на еволуцијата во фамилии на песни преку разлики во нивните варијанти, на збороформи и др. А со создавање на такви дигитални мапи, класификации и подредувања на македонските народни песни би се овозможиле нови бранови на научни истражувања во полето на македонски дигитални хуманистички науки (Digital Humanities) и компјутерски литературни студии (Computational Literary Studies).

ЛИТЕРАТУРА

Кирилични изданија

Бадева, Р. 2012. *Мелодграфски записи на ѝесниѝе на Никола Бадев*. Скопје: Дирекција за култура и уметност.

Бурић, М. 2016. „Интертекстуалност народне епске десетерачке прјесме као лингвостилистички феномен (у свјетлу Вукове теорије)“. Књ. 9: *Девѝти лингвистички скуѝ „Бошковићеви дани“*. Подгорица: Одјељење хуманистичких наука Црногорске академије наука и умјетности, 203–214.

Васиљевић, М. А. 1953. *Југословенски музички фолклор II: Народне мелодије које се ѝевају у Македонији*. Београд: Просвета.

Верковић, С. И. 1860. *Народне ѝесме македонски Буџара*. Књига прва: *женске ѝесме*. Београд: Правителственомъ књигопечатњомъ.

Веселиновиќ, М. В. 1890. „Шар“. *Брасѝво*, бр. 4. Београд: Друштво Св. Саве, 95–130.

Димовска, К. 2014. „Книжевноста како Јанус – за дихотомијата меѓу усната и пишаната книжевност“. *Филолошки сѝудии*, год. 12, бр. 1, 54–70.

Димовска, К. 2022. „Хронотопот и народната мудрост во шест епски песни“. *Македонски фолклор*, год. LIII, бр. 82. Скопје: Институт за фолклор „Марко Цепенков“, 253–274.

Димчевски, Ѓ. 1986. *Сегенка се збира: хармонизирани и обработѝени македонски народни ѝесни*. Скопје: Културно-просветна заедница на Скопје.

Ђурић, М. 2013. „Зашто певач није песник?“. *Токови изучавања српског усменог стваралаштва у другој половини двадесетог века*, Српска књижевна критика друге половине XX века – типолошка проучавања. Београд: Институт за књижевност и уметност.

Јаневска, А. 2024. *Македонското традиционално народно пеење во: Мала Пресја, Голо Брдо и Гора*. Скопје: Институт за фолклор „Марко Цепенков“.

Каракаш, Б. 1995. *Антологија на македонски народни песни*. Скопје: Дега.

Каранов, Е. 1889. „Показалец на напечатаните до сега, в различни сборници и периодически списания, варијанти от песните в Сборника на братя Миладиновци“. *Сборник за Народни Умошворения, Наука и Книжнина – Книга I*. София: Издава Министерството на Народното Просвещение, Държавна печатница, 157–175.

Кузман, А. 2021. „Чалгацијата како создавач/креатор“. *Македонски фолклор*, год. LII, бр. 79. Скопје: Институт за фолклор „Марко Цепенков“, 51–57.

Поленаковиќ, Х. 1952. *Страници од македонската книжевност*. Скопје: „Кочо Рацин“ – издавачко претпријатие за уметничка литература.

Спироски, О. 2025. „Македонски народни песни, варијанти и интертекстуалност“. *Спектар*, бр. 85. Скопје: Институт за македонска литература, 239–265.

Спироски, О., А. Кузман, В. Стојкова-Серафимовска, Д. Даутовски и Љ. Коцарев. 2024. „Класификација на мелодии од народни песни и од црковни напеви од Македонија и од Русија со биоинформатички алгоритми“. *Македонски фолклор*, год. LV, бр. 86. Скопје: Институт за фолклор „Марко Цепенков“, 5–33.

Тончева, В. 2021. „Градско-селскиот тип песни в България от края на XIX и началото на XX век – словесно и музикално съдържание“. *Зборник радова Савремена српска фолклористика IX са Меѓународног научног скупа одржаног од 2. до 4. октобра 2020. у Тршићу, Удружење фолклориста Србије*, Београд. Сикимић, Б., Б. Златковић и М. Думнић Вилотијевић (ур.), 243–265.

Кулавова, К. (прир.). 2021. *ПОИМНИК на книжевната теорија – второ дополнително и изменето издание*. Скопје: Македонска Академија на Науките и Уметностите.

Узунски, Д. 1991. „Запали се, маце, планината“. *Македонски фолклор*, год. XXIV, бр. 47. Скопје: Институт за фолклор „Марко Цепенков“, 133–143.

Фирфов, Ж. (ур.). 1953. *Македонски музички фолклор: Песни I*. Скопје: Книгоиздателство „Кочо Рацин“.

Фирфов, Ж., М. Симоновски и Р. Проданов (ред.). 1959. *Македонски музички фолклор: Песни 2*. Скопје: Фолклорен институт на НР Македонија, Оддел за народна музика и кореографија.

Фирфов, Ж. и М. Симоновски (ред.). 1962. *Македонски мелографи од крајот на XIX век*. Скопје: Институт за фолклор.

Хаџи-Манев, В. (ур.). 1955. *Македонски народни песни 3*. Скопје: Книгоиздателство за уметничка литература „Кочо Рацин“.

Хаџиманов, В. 1960. *Македонски борбени народни песни*. Скопје: Книгоиздателство за уметничка литература „Кочо Рацин“.

Латинични изданија

Bartók, B. 1951. *Serbo-Croatian Folk Songs*. Texts and Transcriptions of Seventy-Five [75] Folk Songs from the Milman Parry Collection and a Morphology of Serbo-Croatian Folk Melodies by Béla Bartók and Albert B. Lord, with a Foreword by George Herzog. New York: Columbia University Press; London: Geoffrey Cumberlege.

Bayard, S. P. 1950. “Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song”. *The Journal of American Folklore*, vol. 63, no. 247, 1–44.

Dawkins, R. 1976. *The selfish gene*. New York: Oxford University Press.

Gopčević, S. 1889. *Makedonien und alt-Serbien*. Wien: Verlag Von L. W. Seidl & Sohn.

Janicki, M., K. Kallio, M. Väina & E. Mäkelä. 2024. “Developing a Digital Research Environment for Finnic Oral Poetry”. *Baltic Journal of Modern Computing*, Vol. 12, No. 4, 535–547.

Jurinić, S. 1892. *Ocjena Djela “Makedonien Und Alt-Serbien” von Spiridion Gopčević*. Sofija: Dvorska tiskara B. Šimáčka.

Kallio, K., M. Janicki, E. Mäkelä & M. Sarv. 2022. “Recognising Intertextuality in the Digital Corpus of Finnic Oral Poetry: Experiment with the Sampo Cycle”. *Digital Humanities in the Nordic and Baltic Countries Publications*, Vol. 4, No. 1, 279–287.

Keller, M. S. 1984. “The Problem of Classification in Folksong Research: a Short History”. *Folklore*, Vol. 95, No. 1, 100–104.

Lévi-Strauss, C. 1974 [1962]. *The Savage Mind*. London: Weidenfeld & Nicolson.

Ljubinković, N. 1991. “Teorija formule A. B. Lorda u svetlu ruralno-urbanih odnosa”. *Македонски фолклор*, год. XXIV, бр. 47. Скопје: Институт за фолклор „Марко Цепенков“, 115–121.

Parry, M. 1971. *The Making of Homeric Verse: The Collected Papers of Milman Parry*. A. Parry (ed.). Oxford: Oxford at the Clarendon Press.

Prechelt, L., G. Malpohl & M. Philippsen. 2002. “Finding Plagiarisms among a Set of Programs with Jplag”. *JUCS – Journal of Universal Computer Science*, vol. 8, No. 11, 1016–1038.

Sağlam T., S. Hahner, L. Schmid & E. Burger. 2024. “Obfuscation-Resilient Software Plagiarism Detection with Jplag”. *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion '24)*. New York: Association for Computing Machinery, 264–265.

Sarv, M. & R. Järv. 2023. “Layers of Folkloric Variation: Computational Explorations of Poetic and Narrative Text Corpora”. *Electronic Journal of Folklore*, vol. 90, 233–265.

Stojkova Serafimovska, V. 2020. “From Love Song to Song of Conflict – Contextual Music Migration”. *Proceedings from the 6th Symposium of the ICTM Study Group for Music and Dance in Southeastern Europe Music and Dance in Southeastern Europe: Migrations, Carnival, Sustainable Development*. Melish, L., N. Green & T. Zebec (eds.). Zagreb: ICTM SG for Music and Dance in SEE – Institute of Ethnology and Folklore Research – ICTM Croatia, National Committee, 111–120.

Suchecka, K. & N. Gasiglia. 2021. “On Digital Comparative Editions and Textual Similarity Detection Tools: Towards a Hypertextual Cartography of a Rewritten Myth”. *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*. Prague: ICL CAS, 163–178.

Сајтографија

“Milman Parry Collection of Oral Literature”, <https://mpc.chs.harvard.edu/> [Пристапено на 23.1.2025].

Chandler, D. 2022. *Semiotics for Beginners*, <http://www.visual-memory.co.uk/daniel/Documents/S4B/sem09.html> [Пристапено на 27.1.2025].

Düring, M., M. Romanello, E. Maud, et al. 2023. “*impresso* Text Reuse at Scale. A Prototype Interface for the Exploration of Text Reuse Data in Semantically Enriched Historical Newspapers”. *Frontiers in Big Data*, Vol. 6, <https://doi.org/10.3389/fdata.2023.1249469> [Пристапено на 4.5.2025].

Foley, J. M. 1988. *The Theory of Oral Composition: History and Methodology*. Bloomington: Indiana University Press, <https://muse.jhu.edu/book/84689> [Пристапено на 22.1.2025].

Janicki, M. 2023. “Large-scale weighted sequence alignment for the study of intertextuality in Finnic oral folk poetry”. *Journal of Data Mining & Digital Humanities*, NLP4DH, <https://doi.org/10.46298/jdmdh.11390> [Пристапено на 4.5.2025].

JPlag v5.1, <https://github.com/jplag/JPlag> [Пристапено на 14.8.2024].

Lord, A. B. 2000 [1960]. *The Singer Of Tales*. Online version of the 2nd edition. Mitchell S. and G. Nagy (eds.). London: Harvard University Press, [The Singer of Tales - The Center for Hellenic Studies](https://www.hcs.harvard.edu/~hellstutj/The%20Singer%20of%20Tales/) [Пристапено на 21.2.2025].

Schöch, C., J. Dudar & E. Fileva (eds.). 2023. *Survey of Methods in Computational Literary Studies* (= CLS INFRA D3.2: Series of Five Short Survey Papers on Methodological Issues), <https://methods.clsinfra.io> [Пристапено на 29.1.2025].

Stalevski, Z. 2005, www.pesna.org [Пристапено на 13.8.2024].

Taylor, A. 1968. "Lists and Classifications of Folksongs". 13. *Jahrgang 1968* (Deutsches Volksliedarchiv. ed). Berlin, Boston: De Gruyter, 1–25, <https://doi.org/10.1515/9783112318096-001> [Пристапено на 4.5.2025].

Линин, А. 1976. *Македонскиот сѝх и најев*. Скопје: Факултет за Музичка и Драмска Уметност, УКИМ [докторска дисертација], <http://hdl.handle.net/20.500.12188/16830> [Пристапено на 31.8.2023].

Мојсова-Чепишевска, В. 2013. „За културната меморија (преку простата и строга македонска песна)“. *Stephanos*, бр. 1. Сетевое издание, Мултиязычный научный журнал, Электронный проект филологического факультета МГУ имени М. В. Ломоносова, 60–77, https://www.stephanos.ru/izd/2013/2013_1_6.pdf [Пристапено на 9.2.2025].

Спироски, М. и О. Спироски. 2009. *Македонски народни ѝесни оо Полог, Македонија: Сардисале Лешочкиот манастѝр, Танец се иѝра срео село и гр. / Macedonian folk songs from Polog, Macedonia: Surrounding of the Lesok Monastery, A dance is played in the middle of the village etc.* Скопје: Спироски М., <http://spiroski.com.mk/книга-2/> [Пристапено на 28.11.2022].

ALGORITHMIC INTERTEXTUALITY FOR CLASSIFICATION OF FOLK SONG VARIANTS AND DETECTION OF SHARED TEXTUAL FORMULAS

Ognen Spiroski

Faculty of Natural Sciences & Mathematics, UKIM – Skopje

Summary

This paper presents the successful use of algorithms for finding the same source code (text) in a collection of program codes (software plagiarism), applied for recognizing sections of the same text in Macedonian folk songs, and for clustering by similarity, finding and classification of textual song variants. This is the first example in the literature of algorithmic intertextuality in oral poetry, folklore and ethnomusicology for the study of song variants, using methods for detecting software plagiarism.

The paper presents a brief review of the literature and research related to intertextuality and variants in folk songs, which has given rise to new classification models and theories in science, as well as a large number of publications with song variants. Algorithmic intertextuality enables new digital analyses of large corpora of songs, variants, and textual formulas.

Two corpora of songs are analyzed with the software tool JPlag v5.1. The first corpus contains 51 textual variants from 6 song families that have already been

studied, and the algorithms cluster them according to their known song families. The second corpus is a large but uncured collection of 1.467 song texts from the website www.pesna.org. The algorithms find and group song variants, as well as a multitude of shared textual formulas between songs, that offer interesting insights into the intertextuality and formulaic nature of folk poetry and its evolution.