

Николче МИЦКОСКИ

Лексикографски центар „Георги Старделов“, МАНУ, Скопје
nmickoski@manu.edu.mk

АНАЛИЗА НА ЛЕКСИЧКАТА ЧЕСТОТА ВО „СИТЕ ЛИЦА НА РЕЧТА“

За потребите на овој труд беше извршена анализа на делото *Ситѝе лица на речѝа* од Катица Кулавакова во издание на Македонската академија на науките и уметностите од 2020 година. Делото претставува збирка на вкупно 58 (педесет и осум) беседи кои акад. Кулавакова ги одржала по разни поводи и беседите структурно се поделени во шест дела.

Во овој труд се користи корпусна лингвистика за да се изврши анализа на лексичката честота. За таа цел, со користење на дигиталната верзија на *Ситѝе лица на речѝа* од Катица Кулавакова (2020) беше создаден корпус. Анализата на создадениот корпус беше извршена со користење на LancsBox v. 5.1.2 наменет за платформата на оперативниот систем Windows (Brezina, V., Weill-Tessier, P., & McEnery, A. 2020). Меѓу научниците постои прифатено мислење дека корпусот претставува збирка на (1) читливи за машини (2) автентични текстови (вклучувајќи и транскрипти од говори) кој е (3) создаден за да биде (4) претставник на одреден јазик или дијалект (McEnery et al. 2006: 5). Исто така, според Reppen (Reppen 2010: 2), „во светот на корпусната лингвистика, корпусот претставува голема, принципиелна колекција на природен јазик што се чува во електронска форма“. Дополнително, според Бибер, Конрад и Репен (Biber, Conrad, and Reppen 1998: 4), корпусната лингвистика претставува истражувачки пристап што ги олеснува емпириските истражувања на јазичните варијации и користење, што овозможува да се добијат резултати од истражувањето што имаат многу поголема генерализација и валидност. Користењето на корпусната лингвистика е избрано затоа што корпусната лингвистика нуди квантитативна и квалитативна перспектива на текстуалните податоци, преку пресметување на лексичката честота односно фреквентност на среќавање и мерење на статистичката

важност, а ги претставува податоците на таков начин на кој истражувачот може да ги анализира индивидуалните среќавања на одредени зборови, квалитативно да ги истражи контекстите во кои тие се среќаваат како колокации, да ги опише забележителните обрасци на среќавање и да ги идентификува функциите во дискурсите (Mautner 2009: 123). За да се добијат податоците за лексичката честота на користењето на зборовите во корпусот што е создаден за овој текст, се користеше карактеристиката за Список на зборови (Words), додека карактеристиката за совпаѓање и анализа на клучни зборови во контекст (Key Word In Context – KWIC) се користеше за да се добијат податоци за користење на термините во различни контексти, односно како секој збор се анализираше како клучен збор во контекст.

Анализата на лексичката честота во создадениот корпус покажува дека во овие две дела има вкупно 75.880 зборови од кои 14.229 се различни зборови, што значи дека речиси 19 % од зборовниот фонд се сосема различни зборови.

Интересно е што дури 60 % од вкупниот број различни зборови, односно 8342 зборови се среќаваат само по еднаш, односно воопшто не се повторуваат. Најголем дел од овие зборови се полнозначни зборови, односно именки, глаголи, придавки и др.

Под термините е даден и графички приказ на тоа со кои зборови најчесто се среќаваат обработените термини, но само ако бројот на зборовите кои се среќаваат во комбинација со обработениот термин е помал од 250, бидејќи графичкиот приказ на број на зборови што е поголем од 250 е неразбирлив. Во центарот на графичкиот приказ се наоѓа терминот што се обработува, а околу него се зборовите кои се среќаваат во комбинација со него и создаваат колокации. Колку поблиску е некој збор до централниот термин, толку почесто се среќава во комбинација со него, а колку е подалеку, толку поретко создава колокација со тој термин. Од левата страна се наоѓаат зборовите што се среќаваат пред централниот термин, додека од десната страна се наоѓаат зборовите што се среќаваат по централниот термин.

Најзастапен збор во корпусот е предлогот „на“ кој се среќава 5.489 пати. Овој предлог се среќава во комбинација со 480 други зборови. Пред него најчесто се среќаваат зборовите „и“, „поезијата“, „јазикот“ и „историјата“ и со нив ги создава колокациите „и на“ којашто се среќава 344 пати, „поезијата на“ којашто се среќава 114 пати, „јазикот на“ којашто се среќава 58 пати и „историјата на“ којашто се среќава 45 пати. По предлогот „на“ најчесто се среќаваат зборовите „македонската“, „светот“, „македонскиот“, „еден“ и „јазикот“. Така, најзастапени се колокациите „на македонската“ којашто се среќава 130 пати, „на светот“ којашто се среќава 101 пат, „на македонскиот“, којашто се среќава 97 пати, „на еден“ којашто се среќава 67 пати и „на јазикот“ којашто се среќава 58 пати.

Во табелата подолу, се претставени десетте најзастапени зборови во корпусот.

Збор	Лексичка честота во корпусот
на	5.489
и	4.390
во	1.943
се	1.729
да	1.360
за	1.189
од	1.186
е	1.164
со	749
како	601

Табела 1: Лексичка честота на зборовите во корпусот

Глагол	Лексичка честота во корпусот
е	1.164
има	411
може	235
биде	174
треба	157
нема	75
постои	75
значи	68
стане	53
претставува	38

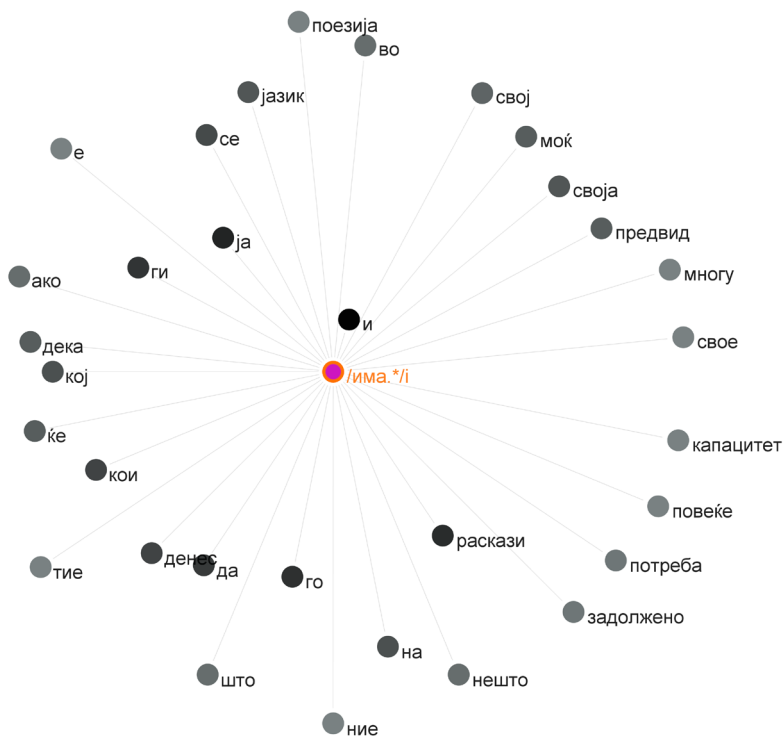
Табела 2: Лексичка честота на глаголите во корпусот

Најзастапен глагол во корпусот е глаголот „е“ којшто се среќава 1.164 пати во корпусот и создава комбинации со 69 различни зборови. Пред овој глагол најчесто се појавуваат зборовите „не“ и тоа 102 пати, „тоа“ 59 пати и „што“ 42 пати, а по него најчесто се појавуваат зборовите „да“ 34 пати, како и „еден“ и „и“ по 21 пат.



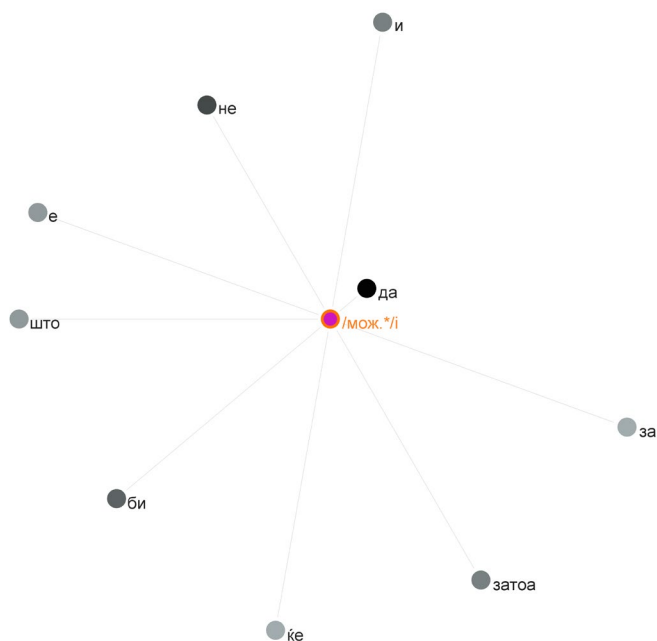
Графички приказ на колокациите со глаголот „е“

Втор најзастапен глагол е глаголот „има“ којшто се среќава 411 пати во корпусот, а создава колокации со 32 други зборови. Најчесто се среќава во комбинација со сврзникот „и“ којшто се појавува 30 пати пред глаголот „има“ и кратката заменска форма „ја“ којашто се појавува 20 пати по глаголот „има“.



Графички приказ на колокациите со глаголот „има“

Трет глагол што најчесто се среќава во корпусот е глаголот „може“ којшто се појавува 235 пати. Најчесто е во комбинација со зборовите „не“ којшто се појавува 52 пати пред него и со „да“ којшто се појавува 188 пати по него, а создава комбинации со 9 различни зборови.



Графички приказ на колокациите со глаголот „може“

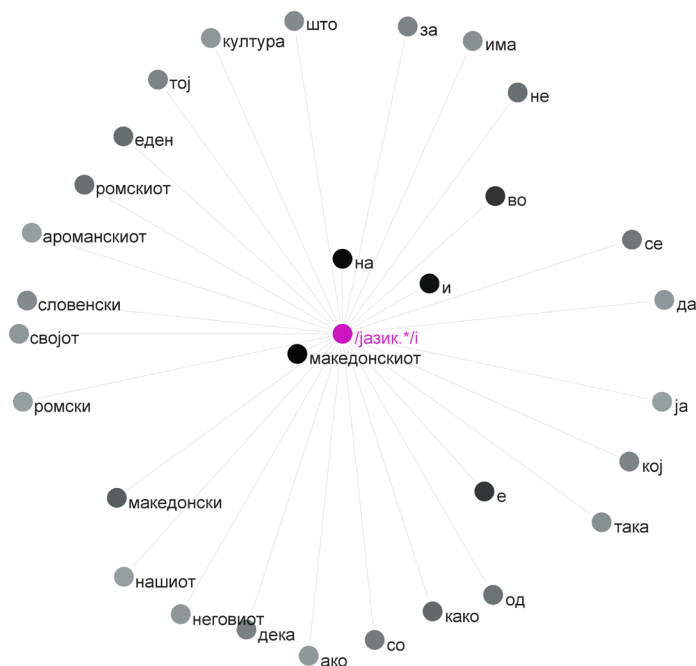
Во табелата 2 подолу е дадена лексичката честота на именките во корпусот.

Именка	Лексичка честота во корпусот
јазик	537
свет	340
поезија	317
народ	223
година	209
култура	202
историја	198
Македонија	167
книжевност	158
време	140

Табела 3: Лексичка честота на именките во корпусот

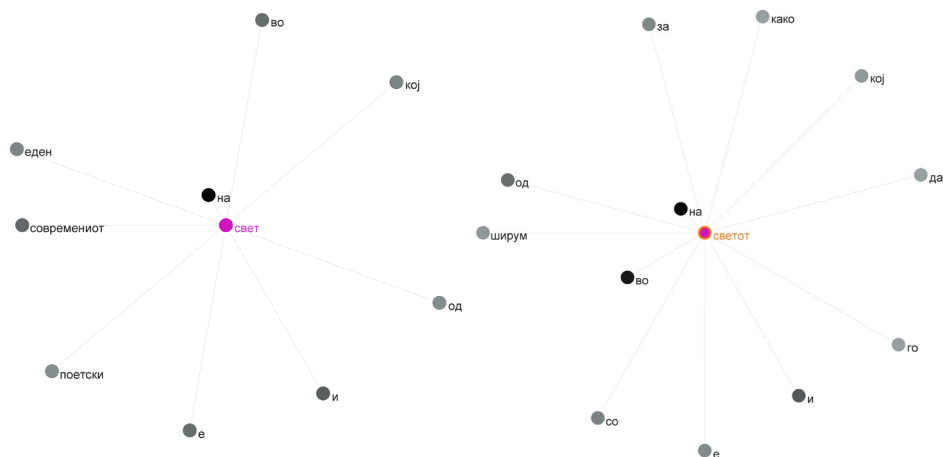
Најзастапена именка во корпусот е именката „јазик“ којашто се среќава 326 пати во основната форма или вкупно 537 пати во разни форми, односно 115 пати во членуваната форма „јазикот“, 74 пати во множинската форма „јазици“ и 19 пати во членуваната множинска форма „јазиците“, како и по еднаш во формите „јазиковед“ и „надјазикот“.

Именката „јазик“ во корпусот се среќава во комбинација со 30 други зборови, а најчесто со придавката „македонскиот“ којашто се јавува 92 пати пред именката и со неа ја формираат колокацијата „македонскиот јазик“, а нечленуваната форма „македонски јазик“ се јавува 20 пати.



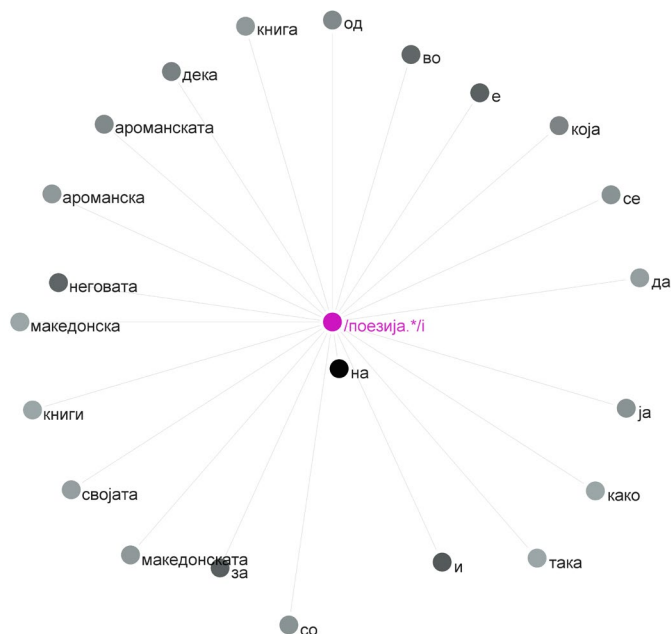
Графички приказ на колокациите со формите на именката „јазик“

Следна најзастапена именка во корпусот по именката јазик е именката „свет“ којашто се среќава 332 пати и тоа 214 пати во членуваната форма „светот“, 118 пати во основната форма „свет“, 7 пати во множинската форма „светови“ и еднаш во членуваната форма „светов“. Формата „свет“ се среќава во комбинација со 9 други зборови и тоа најчесто во колокациите „на свет“ 46 пати и „современиот свет“ 10 пати. Членуваната форма „светот“ создава колокации со 12 други зборови и тоа најчесто со предлозите „на“ и „во“ со кои ги формира колокациите „на светот“ односно „во“ коишто се среќаваат 101 пат односно 73 пати во корпусот.



Графички приказ на колокациите со формите на именката „свет“

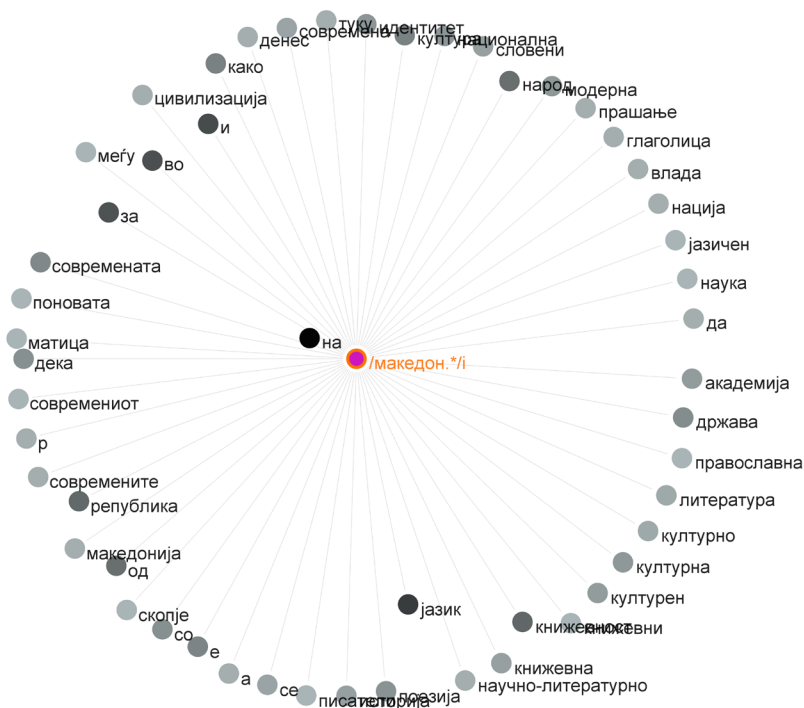
Потоа следува именката „поезија“ којашто се среќава 317 пати и тоа 162 пати во основната форма „поезија“ и 155 пати во членуваната форма „поезијата“. Оваа именка во двете свои форми создава колокации со 22 други зборови, од кои најчесто со предлогот „на“ кој се појавува 131 пат пред неа.



Графички приказ на колокациите со формите на именката „поезија“

Анализата на корпусот ја потврди констатацијата на Бибер дека именките се користат многу повеќе од заменките во академските излагања (Biber et al. 1999: 235).

Од придавките, најчесто, дури 686 пати се среќаваат придавките изведени од именката Македонија и тоа: „македонската“ 230 пати, „македонскиот“ 198 пати, „македонски“ 93 пати, „македонска“ 63 пати, „македонското“ 46 пати, „македонските“ 44 пати, „македонско“ 11 пати и „македонскиов“ еднаш. Од анализата може да се забележи дека зборот „Македонија“ е еден од најпродуктивните зборови во овој корпус. Се појавува во дури 38 различни форми, а тие различни форми се среќаваат 930 пати и создаваат колокации со 53 други зборови во корпусот, а од полнозначните зборови, најчесто се среќаваат во комбинација со именките „јазик“ 112 пати, „книжевност“ 40 пати, „република“ 39 пати и „народ“ 33 пати.



Графички приказ на колокациите со формите на зборот „Македонија“

Покрај придавките изведени од именката Македонија, има голема застапеност и на заменските придавки, од кои најчести се „неговата“ која се среќава 123 пати, „својата“ која се среќава 90 пати и „неговиот“ која се среќава 77 пати.

Од заменките, многу почесто се среќаваат кратките заменски форми „го“ 471 пати, „ја“ 389 пати, „ги“ 297 пати, а поретко заменките „тоа“ 303 пати, „тој“ 313 пати, „кој“ 211 итн.

Покрај овие анализи, беше направена анализа на конструкциите што најчесто се користат во овој корпус. Во анализата се опфатени конструкциите што се составени од два и три члена.

Како што може да се види од табелата подолу, во конструкциите со два члена, половината, односно пет од нив се да-конструкции.

Конструкција со два члена	Лексичка честота во корпусот
да се	509
и на	342
и во	147
и да	131
на македонската	130
треба да	119
да го	107
и за	103
може да	102

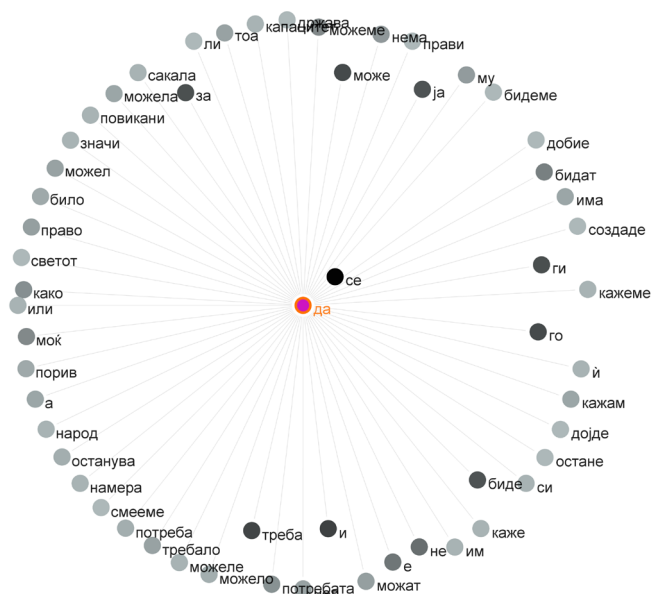
Табела 4: Лексичка честота на конструкциите со два члена во корпусот

Табелата подолу покажува дека и во конструкциите со три члена преовладуваат да-конструкциите. Од десетте најчести конструкции со три члена, првите четири се да-конструкции.

Конструкција со три члена	Лексичка честота во корпусот
треба да се	63
може да се	63
за да се	42
и да се	39
на Ацо Шопов	36
на македонскиот јазик	36
македонскиот јазик и	24
на македонскиот народ	23
поезијата на Ацо	22
на меѓународниот ПЕН	21

Табела 5: Лексичка честота на конструкциите со три члена во корпусот

Зборот „да“ во корпусот се јавува како сврзник и како честичка. Тој е петти најкористен збор во овој корпус и се среќава вкупно 1.360 пати. Сепак, тој е најпродуктивен кога станува збор за создавање конструкции. Најчесто се среќава во комбинација со „и“, „треба“ и „може“ кои се јавуваат пред него, како и „се“, „го“ и „ги“ кои се јавуваат по него.



Графички приказ на колокациите со „да“

Користењето на корпусната лингвистика за да се анализира корпусот создаден од беседите на Кулавакова збрани во „Сите лица на речта“ ни овозможи да продере подлабоко во зборовниот фонд што Кулавакова го користи во своите беседи. Анализата со користење на корпусната лингвистика во овој труд ни овозможи да добиеме многу корисни информации во врска со афинитетите и тенденциите на Кулавакова во поглед на лексиката што ја користи, конструкциите и колокациите што најчесто ги користи и слично. Ни овозможи да имаме подобар увид во нејзиниот автентичен јазик што го користи во реално време, зборовите, но и фразите што најчесто ги користи. Можеби уште поважни се информациите во врска со зборовите што ретко се користат затоа што на тој начин може да се збогати зборовниот фонд на македонскиот јазик со нови зборови. Така, забележуваме појава на зборови што многу ретко се користат во македонскиот јазик, а некои од нив не се вклучени и опишани во толковните речници на македонскиот јазик. Такви зборови се идеологема, јазиковед, топос, мизансцен, мапштаб, светотвори (светотворна) и други.

Оваа анализа претставува само загревање на површинскиот слој од слоевитата текстуалност и интертекстуалност што ни ја нуди творештвото на Катица Кулавакова. На крајот, може да заклучиме дека користењето на корпусната лингвистика ќе ни отвори широко поле за понатамошни истражувања на творештвото на Кулавакова, но и на македонското творештво воопшто.

Библиографија

- Ќулавкова, К. 2020. *Сите лица на речта*. Скопје: Македонска академија на науките и уметностите
- Biber, D., Conrad, S., Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited
- Brezina, V., Weill-Tessier, P., McEnery, T. 2020. #LancsBox v. 5.1.2 [software]
- Mautner, G. 2009. Checks and balances: how corpus linguistics can contribute to CDA. In *Methods of Critical Discourse Analysis*, eds. Wodak P., Meyer, M., pp. 122-144. London: Sage
- McEnery, T., Xiao R., Tono, Y. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. New York: Routledge
- Reppen, R. 2010. *Using corpora in the language classroom (Cambridge language education)*. New York: Cambridge University Press

Nikolche MICKOSKI

ANALYSIS OF THE LEXICAL FREQUENCY IN *SITE LICA NA RECHTA*

The paper “Analysis of the lexical frequency in *Site lica na rechta*” by Nikolche Mickoski gives an overview of the lexical frequency and key word in context analysis of the corpus created from *Site lica na rechta* by Katica Kjulavkova. The use of corpus linguistics provides useful information about the lexis, constructions and collocations used in the corpus, and gives an insight into the use of the author’s authentic language. The paper provides an overview of the most frequent words, verbs, nouns, adjectives and pronouns with statistical data and graphical representations of the collocations they create. The analysis sheds light on the words used by Kjulavkova that are rarely used in Macedonian.

Keywords: Corpus linguistics, lexical frequency, key word in context, collocations, speech, Macedonian language

