811.163.41'322
811.163.41'27:|004.773:621.395.721.5
https://doi.org/10.18485/msc.2018.47.3.ch10

Бранислава Б. ШАНДРИХ
Филолошки факултет
Универзитета у Београду

Душко М. ВИТАС
Природно-математички факултет
Универзитета у Београду

КВАНТИТАТИВНИ ПРЕГЛЕД ЈЕЗИКА КРАТКИХ ПОРУКА

У раду је направљена статистичка анализа корпуса СМС порука. Корпус чини приближно осам хиљада кратких порука размењиваних између учесника различитог узраста у четворогодишњем периоду. На оригиналне поруке првобитно је примењен поступак обраде, након чега је корпус обрађен помоћу језичког алата *Јунитекс*. Групе препознатих и непрепознатих токена[1] разврстане су по различитим категоријама, а затим је анализирана њихова расподела. Узимајући добијене вредности у обзир, изведен је закључак да и језик СМС порука тежи ка што већој сличности са говорним језиком, као што је већ раније утврђено за поруке социјалне мреже Твитер.

Кључне речи: српски језик, нестандардни језик, СМС поруке, кратке поруке, корпусна лингвистика

1. Увод

Годинама уназад, писана комуникација обавља се посредством мобилних телефона, рачунара, таблета и сличних уређаја. Посредници у комуникацији су разноврсни: од СМС порука, до разних социјалних мрежа (Фејсбук, Твитер, Инстаграм), апликација за размену кратких порука (Вајбер, Вотсап, Телеграм), ћаскаоница итд. У овом раду посебно се осврћемо на комуникацију СМС порукама. Специфичност овакве размене порука лежи у начину њиховог стварања, ограничењима које оно са собом носи, као и у њиховој намени. СМС поруке размењују се употребом мобилних телефона, што под-

[1] *Токен* је низ било каквих карактера који нису сепаратори (размак, нови ред, табулатор).

разумева унос тридесетак карактера који су распоређени на малом простору екрана. Овакав унос текста носи са собом и ризик прављења одређених врста грешака, а нарочито да се жељено слово замени њему суседним. Друга специфичност ове комуникације лежи у њеним ограничењима. Наиме, дужина једне поруке ограничена је на сто шездесет карактера, што обухвата и размаке и знаке интерпункције. Прва последица овог ограничења јесте то што аутор поруке, свестан ограничења, тежи ка свесном сажимању садржаја. Занимљиво је приметити и то да је задржан тренд просечне дужине поруке, у оквиру задатих ограничења, иако многи корисници мобилних телефона данас имају на располагању велики број СМС порука захваљујући конфигурацији постпејд пакета који користе. Претпостављамо да је ово последица управо непрактичних модерних тастатура за унос карактера, а тврдњу поткрепљујемо емпиријским резултатима анализе садржаја порука.

Друго ограничење односи се на употребу латиничних дијакритика, као и ћирилице. У порукама које у себи садрже бар један дијакритик или ћирилично слово, ограничење дужине смањено је на само 70 карактера. Последица је распрострањена употреба тзв. „ошишане латинице" и готово потпуно одсуство ћирилице у СМС порукама.

Намена СМС порука је да кратко и концизно саопште или затраже информацију. Ово није случај код порука других посредника у комуникацији (нпр. код социјалних мрежа уобичајне су јавне објаве мишљења), што представља још једну специфичност комуникације СМС порукама. Проблеми категоризације порука на основу њиховог садржаја (енг. *topic classification*) и анализе осећања (енг. *sentiment analysis*), који су иначе углавном успешно решавани над дигитализованим текстом, због поменутих ограничења представљају својеврстан изазов.

Ради уштеде времена и новца, аутори СМС порука теже ка скраћивању речи. Још један феномен који прати кратке поруке, последица је немогућности изражавања емоција посредством саме писане комуникације. Аутори порука тада проналазе различите начине да примаоцу поруке приближе своје расположење: употребом емотограма (енг. emoticons), великих слова за наглашавање, понављаном интерпункцијом или пригодним скраћеницама. Према Турлоу 2003, текст СМС порука је ближи говорном него писаном језику. Занимљиво је да, упркос свим ограничењима, различити начини којим се она превазилазе не утичу на разумљивост порука. Детаљна анализа стандардних грешака које праве корисници при размени кратких порука направљена је у Миличевић 2017.

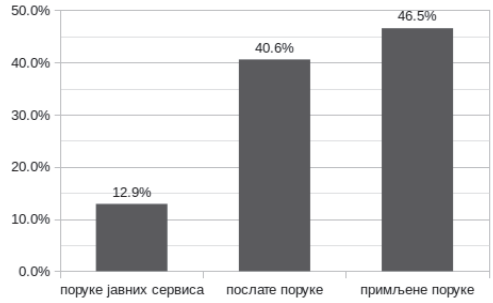
У даљем излагању ћемо у другом одељку описати поступак припреме корпуса прикупљених СМС порука, док у трећем одељку правимо детаљан квантитативни преглед резултата добијених обрадом алатом *Unitex* (даље *Јунитекс*). У четвртом одељку дискутујемо резултате, док у петом одељку истичемо најважније закључке спроведене анализе.

2. Припрема корпуса

Кратке поруке анализиране у овом раду размењиване су између неколико десетина различитих учесника према полу и узрасту. Доминантни језик порука је српски, али постоје и поруке на енглеском и немачком. Писмо порука је углавном латиница, али има и ћириличних слова. Од укупно 7.855 порука, 1.013 (приближно 12,9%) чине аутоматски генерисане поруке јавних сервиса: паркинг, банка, извештаји такси служби и слично. Преостале 6.842 поруке (приближно 87,1%) састоје се од 3.186 (приближно 40,6%) послатих и 3.656 примљених порука (приближно 46,5%, видети Графикон 1).

Пре пропуштања корпуса кроз *Јунитекс*,[2] на поруке је примењено неколико корака препроцесирања. Полазни корпус састоји се од порука обележених у XML формату,[3] при чему једна порука у склопу атрибута своје етикете носи различите информације о себи: број телефона примаоца, односно пошиљаоца, да ли је порука примљена или послата, датум и време слања поруке, да ли је порука успешно испоручена, као и сâм текст поруке:

```
<sms address="+3816****" type="0" body="Ne treba" read="1" date="1449247643732"/>
```



Графикон 1. **Расподела кратких порука према извору**

Први корак јесте елиминација свих вредности, осим садржаја поруке. Након тога, следе кораци „чишћења". Према Витас 1979, потребно је прво сваки дијакритик српске латинице заменити својим еквивалентом (č → cy, ć → cx, đ → dx, ž → zx, š → sx, dž → dy, lj → lx, nj → nx, слично и за одговарајућа велика слова). С обзиром на то да корпус садржи и поруке на немачком језику, без губитка информација, обављена је и замена умлаута одговарајућим еквивалентима (ü → ue, ö → oe, ä → ae, ß→ ss). Слично је и са ћирилицом, али су у овом случају сви ћирилични карактери замењени токеном *constcyr*. Садржај ових порука није преведен у латинично писмо како би се задржала информација о употреби ћирилице у конверзацији СМС порукама. Како XML не дозвољава појаву одређених карактера у садржају (&,

[2] Unitex, http://unitexgramlab.org/
[3] XML, https://www.w3.org/XML/

<, >, „, и ’), да не би били изгубљени, чувају се у тексту у виду одговарајућих карактерских ентитета (нпр. уместо < стоји *<*). Ти карактерски ентитети замењени су својим одговарајућим, оригиналним карактером.

Међу последњим корацима припреме корпуса јесте и канонизација емотограма. Како се емотограми попут „:-)” и „:-(” не би, приликом обраде *Јунитексом*, рашчланили и посматрали као засебни интерпункцијски знаци, различите групе емотограма замењене су одговарајућим токенима (видети Табелу 1).

Група	**Токени**
emotclosedhappy	xD x-D xDDD x-DDD x) x))) x-) x-)))
emothappy	:D :-D :DDD :-DDD
emotkiss	:* :-* :*** :-***
emotheart	<3 <333
emottongue	:P :-P :PPP :-PPP :p :-p :ppp :-ppp
emotsad	:(:-(:(((:-(((:-[:[:-[[[:[[[
emotconfused	:/ :-/ :/// :-///

Табела 1. Групе емотограма

На крају, обављено је и груписање одређених карактера према ономе што та целина представља. На пример, све електронске адресе замењене су са *constemail*. За остале замене, погледати Табелу 2.

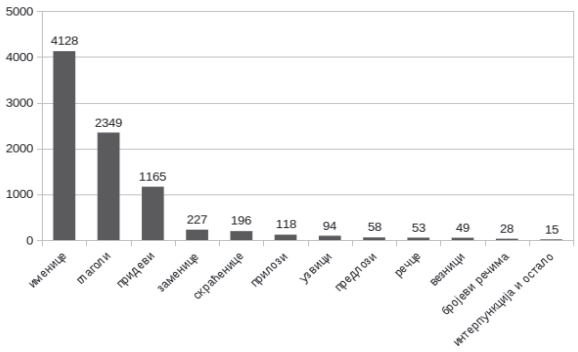
Група	**Објашњење**
constpercent	Вредност у процентима
constmoney	Некакав новчани износ
constphonenumber	Број телефона корисника
constservicenumber	Број телефона сервиса
constregplates	Број регистарских таблица
consttime	Време
constlongnumber	Група више узастопних цифара које нису препознате као целина
constdigit	Цифра која није припала ниједној поменутој категорији

Табела 2. Групе емотограма

3. Резултати обраде *Јунитексом*

Након припреме на начин описан у претходном одељку, корпус је спреман за обраду. У свим кратким порукама препознато је укупно 194.975 токена, односно 11.224 различитих токена. Међу њима је 2.744 (приближно 24,5%) токена који не постоје у *Јунитексовим* електронским речницима (Крстев 2006; Витас 2008; Крстев 2010). Међу препознатим токенима, препознато је и 158 група од више токена које представљају полилексичке јединице, као што су: аутобуска станица, крвна слика, Црна Гора, благо теби, добро јутро итд. Просечан број токена по поруци је приближно 14, а просечна дужина поруке је приближно 81 карактер (отприлике 50%, у односу на горње ограничење дужине од 160 карактера).

Речи присутне у електронским речницима аутоматски су обележене категоријом речи којој припадају (видети Графикон 2).[4]

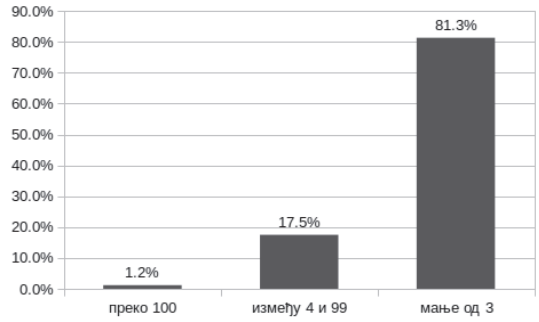


Графикон 2. **Расподела препознатих речи**

Препознато је највише именица (4.128 различитих), потом глагола (2.349 различитих) и придева (1.165 различитих), а најмање има различитих речи које изражавају некакву количину (попут „три", „обе", „једном", 28 различитих) и знакова интерпункције и осталих помоћних симбола (запете, заграде, тачке и слично).

Према броју појављивања, најучесталији токен је запета, затим следе везници и заменице, потом нумеричке вредности и датуми. Са фреквенцијом већом од сто има 132 токена (приближно 1,2%), са фреквенцијом између четири и деведесет и девет има 1.963 токена (приближно 17,5%), док токена са фреквенцијом мањом или једнаком три има 9.129 (приближно 81,3% од 11.224 различитих токена, видети Графикон 3).

[4] Не говори се о граматичком концепту врсте речи, већ о категоријама које се речима доделе приликом уноса у електронски речник.

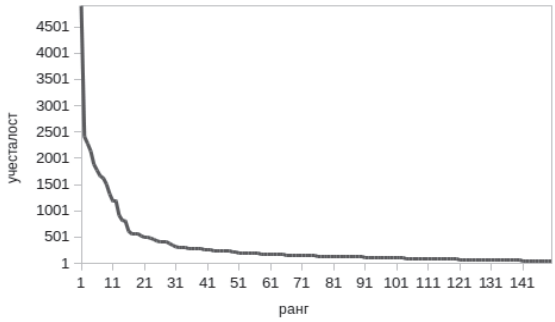


Графикон 3. **Учесталост токена**

На Графикону 4 приказан је однос најучесталијих токена и њихове заступљености у корпусу. Токени су прво рангирани према учесталости, опадајуће: најчешћи токен (у овом случају запета) има ранг један, затим следећи (са нешто мањом учесталошћу) ранг 2 итд. Ранг је приказан на *x*-оси. На *y*-оси представљен је број појављивања. Одатле се може закључити да овај корпус прати Зипфову расподелу: токена високе учесталости има мало, али они чине велики део корпуса; с друге стране, велики је број токена чије фреквенције чине готово занемарљиво мали део корпуса.

Не би било сасвим прецизно закључити да је оваква дистрибуција токена последица искључиво богате лексике СМС порука. Више о лексици кратких порука може се закључити детаљнијом анализом непрепознатих токена.

Непрепознати токени ручно су разврстани у 12 различитих категорија. Ако је токен додељен некој категорији, то значи да токен потенцијано није препознат електронским речником због онога што та категорија представља. Преглед категорија и објашњења дат је у Табели 3.



Графикон 4. **Зипфов закон**

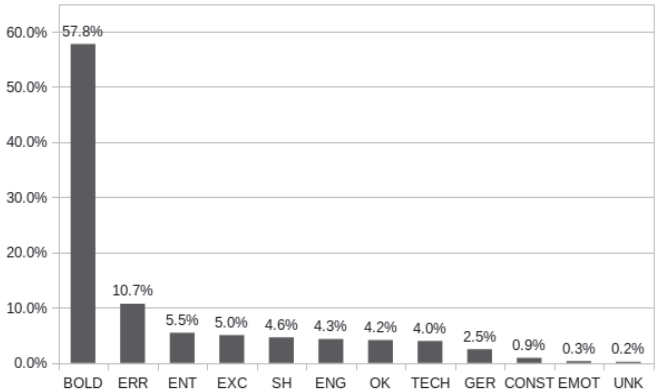
Једна реч би могла припадати и у више различитих категорија. На пример, реч „mpze” има и грешку при куцању, али је написана и „ошишаном” латиницом. Реч је ипак додељена категорији неисправно откуцаних речи, јер је то грешка коју прво треба елиминисати у случају некакве друге, дубље

анализе. Расподела непрепознатих речи према категоријама приказана је на Графикону 5.

Категорија	Објашњење	Пример
BOLD	„Ошишана" латиница	poslacu, racunar, secer
ERR	Погрешно написан појам	kontropni, tvpj, mpzda
ENT	Властита именица, малим словом	ivana, grigor, dunav
EXC	Узвик	ajjjjj, eejjjj, eo, bree
SH	Скраћеница	vrv, np, nnc, npm, stv
ENG	Реч на енглеском језику	voice, year, really,
OK	Исправно написан појам	odlazni, propolis, latino
TECH	Технички појам	www, wifi, css, dropbox
GER	Реч на немачком језику	arbeite, dich, frei
CONST	Предефинисане групе карактера	consttime, constdigit
EMOT	Емотограми	emotsad, emotkiss
UNK	Ниједна од постојећих категорија	xxx, zl

Табела 3. Категорије непрепознатих токена електронским речником

На Графикону 5 може се видети да је приближно 4,6% (127) речи непрепознато због тога што представљају некакав скраћени облик. У Табели 4 приказујемо неке од често коришћених скраћеница, са њиховим значењима, сортирано према заступљености у корпусу.



Графикон 5. **Расподела непрепознатих речи**

Скраћено	Значење	Скраћено	Значење
ok	океј	bzvz	безвезе
zvrc	зврцнути	nzm	не знам
odg	одговорити	tnx	енг. ”thanks”
pozz	поздрав	zab	заборавити
vcs	вечерас	ustv	у ствари
fb	фејсбук	vs	видимо се
msm	мислим	nmp	немам појма
vrv	вероватно	np	нема проблема
najvrv	највероватније	stv	стварно
nmg	не могу	otp	отприлике
pls	енг. “please”	ozb	озбиљно

Табела 4. СМС лексика

Међу непрепознатим речима, готово једнако заступљена је и категорија узвика (Графикон 5). У Табели 5 приказане су речи додељене овој категорији, са интерпретацијама аутора, сортирано према заступљености у корпусу. Ради се о речима у којима се најчешће вокал (самогласник) понавља више пута. Наглашавањем речи на овај начин, аутор поруке жели да нагласи тон поруке: усхићење, захвалност, тугу, опомињање и слично. У Шандрих 2018б, управо полазећи од ове претпоставке, речи са поновљеним узастопним карактерима улазе у поступак анализе осећања у кратким порукама. Још једна од карактеристика текста која је ушла у поменуто истраживање јесте и истицање тона речи употребом великих слова, као и понављањем интерпункције (*Gde si?* → *GDE SI???*). Овај рад није посвећен таквој врсти анализе.

Скраћено	Значење	Скраћено	Значење
weeeeee	радовање	finoo	наглашавање
ae	„хајде”	hahaha	смејање
ajjjjj	„хајде”	хахаха	смејање
ajoj	нелагода	hihi	осмехивање
auh	нелагода	hejj	дозивање
bozeeee	ишчуђавање	jeeej	радовање
bree	наглашавање	juuu	ишчуђавање

ccc	цоктање	mhm	потврђивање
dobrooo	наглашавање	neee	негирање
ejjjj	дозивање	vauuu	одушевљење
jook	наглашавање	mmmm	размишљање

Табела 5. СМС лексика

4. Дискусија

Управо због свих својих специфичности, СМС поруке могу послужити као основа различитим истраживањима. Већ због ограничене дужине, носе много мање информација у односу на некакав дужи текст (новински чланак, роман и слично). Самим тим захтевају посебан начин приступа решавању иначе, над дужим текстовима, добро решеним проблемима (попут категоризације на основу теме, анализе осећања итд.). Идентификација пошиљаоца СМС поруке на основу специфичности приликом куцања предложена је у Шандрих 2018а. Анализа осећања у СМС порукама на основу емотограма, скраћеница и стилометричких карактеристика предложена је у Шандрих 2018б. У оба рада, значење речи употребљених у текстовима порука није учестововало у анализи. На пример, на текст поруке могла је бити примењена анотација врсте речи (енгл. *Part-of-Speech Tagging*). Ипак, као што је објашњено у претходном поглављу, известан број речи, из различитих разлога, није препознат електронским речницима. На пример, према Графикону 5, приближно 57,8% речи није препознато због тога што је уместо латиничних дијакритика коришћена тзв. „ошишана" латиница. Проблем рестаурације дијакритика у српском разматран је у Крстев 2018 и предложено решење могло би бити примењено на корпус кратких порука у циљу редукције речницима непрепознатих токена.

Аутори у Миличевић 2017 детаљно разматрају различите врсте грешака које аутори кратких порука уобичајно праве. Пратећи резултате до којих су аутори дошли, може се направити обрнути поступак корекције грешака како би се добили исправни облици.

5. Закључак

Оно што се на основу статистичких резултата над датим корпусом може закључити јесте потврда о томе да аутори кратких порука на различите начине теже ка асимилацији писаног текста са говором: наглашавањем речи употребом великих односно понављаних слова, понављањем интерпункције, употребом емотограма и различитих скраћеница којима поруке добијају неформалан тон итд. Упркос свим овим појавама, разумљивост

порука није угрожена, а још једна од последица јесте и развијање нове лексике СМС порука.

ЛИТЕРАТУРА

Витас 1979: Duško Vitas, „Prikaz jednog jistema za automatsku obradu teksta", у: *Simpozijum INFORMATICA, Bled*, 7–10.

Витас 2008: Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. Tour du Monde through the Dictionaries. In M. Constant, T, Nakamura (eds.) *Actes Du 27eme Colloque International Sur Le Lexique et La Gammaire*, L'Aquila, Universite Paris-Est, 249–256.

Крстев 2006: Krstev, Cvetana, Duško Vitas, and Agata Savary. Prerequisites for a Comprehensive Dictionary of Serbian Compounds." In *FinTAL 2006*, LNAI 4139, Springer, 552–563.

Крстев 2010: Krstev, Cvetana, Ranka Stanković, Ivan Obradović, Duško Vitas, and Miloš Utvić. 2010. Automatic Construction of a Morphological Dictionary of Multi-Word Units. In *IceTAL 2010*, LNAI 6233, Springer, 226–237.

Крстев 2018: Krstev, Cvetana, Ranka Stanković, and Duško Vitas. Knowledge and Rule-Based Diacritic Restoration in Serbian. In *3rd International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, 104–114.

Миличевић 2017: Maja Miličević Petrović, Nikola Ljubešić i Darja Fišer, „Nestandardno zapisivanje srpskog jezika na Tviteru: mnogo buke oko malo odstupanja?" *Анали Филолошког факултета*, 29 (2), 111–136.
https://doi.org/10.18485/analiff.2017.29.2.8.

Турлоу 2003: Thurlow, Crispin, and Alex Brown. Generation Txt? The Sociolinguistics of Young People's Text-Messaging. *Discourse Analysis Online* 1 (1) https://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html

Шандрих 2018а: Šandrih, Branislava. 2018a. Fingerprints in SMS Messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting. In *3rd International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, pp. 203–210. Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Шандрих 2018б: Šandrih, Branislava. SMS Sentiment Classification Based on Emoticons, Informal Abbreviations and Other Text Features. In *International Quantitative Linguistics Conference QUALICO 2018*. Wroclaw, Poland: Institute of Library and Information Science / Faculty of Mathematics and Computer Science (University of Wroclaw).

Branislava Šandrih, Duško M. Vitas

LANGUAGE OF SMS MESSAGES: A QUANTITATIVE OVERVIEW

Summary

This paper analyzes the corpus of SMS messages. The corpus makes about eight thousand short messages exchanged between participants of different ages in a four-year period. The original corpus was firstly cleaned, after which the corpus was processed using the Unitex language tool. Groups of recognized and unrecognized tokens were classified in different categories, and then their distribution was analyzed. By taking the acquired values into consideration, it is concluded that the language of the SMS message strives for the greater similarity with the spoken language, which has already been established for the messages of the social network Twitter.