81'322:061.1EU
Оригинални научни рад

Наслеђе 27 • 2014 • 139-154

Игор Д. Ивановић[1]
Институт за стране језике
Подгорица

КОРПУСНА И РАЧУНАРСКА ЛИНГВИСТИКА НА ПРИМЈЕРУ УПИТНИКА ЕК[2]

Овај рад се бави анализом дијела Упитника Европске комисије. Из цјелокупне анализе, која је изложена у нашем докторском раду, за потребе овог рада смо одабрали два аспекта. Први је анализа имперсонализованог регистра, а други представља налажење везе између *hapax legomena* и грешака које смо уочили у поменутом корпусу. Показаћемо како употреба корпуса и рачунара може да баци ново свјетло на традиционална лингвистичка истраживања. Поред тога, овај рад треба схватити као својеврсну промоцију корпусне и рачунарске лингвистике које тек треба да нађу своју пуну примјену у оквиру лингвистичких истраживања.

Кључне ријечи: корпус, рачунари, лингвистичка анализа, Упитник ЕК

Анализа корпуса

Увод

Ова анализа настала је као производ рада на корпусу Упитника ЕК са којим смо дошли у контакт током превођења истог. Захваљујући овом, били смо у стању да прикупимо велику збирку текстова из различитих области како бисмо могли да задовољимо критеријум репрезентативности корпуса (Biber 1988: 56-63). Као што смо навели, одабрали смо два аспекта: имперсонализовани регистар и однос између *hapax legomena* и грешака које смо уочили. Први аспекат смо одабрали зато што је у српском језику имперсонализовани регистар обично био анализиран на много мањим корпусима. Други аспекат смо одабрали како би показали да интуитивни приступ анализи корпуса може бити веома погрешан, као и да налажење и анализа правописних грешака може довести до бољих рачунарских програма који се баве исправљањем текстова.

1 iggybosnia@gmail.com

2 Овај рад представља дио нашег истраживања за потребе докторског рада: „Контрастивна анализа стручне терминологије докумената ЕУ кроз перспективу корпусне лингвистике".

Методологија истраживања

Методологија на општем плану

Овај рад је замишљен као рад у којем се, на општем плану, прожимају двије методологије: контрастивна и корпусна методологија рада. Прву методологију смо користили јер смо кренули од основне поставке да се контрастирањем разлика и сличности двају језика (српског и енглеског) могу анализирати одвојени или заједнички морфолошки, синтаксички, семантички или лексички елементи. Ова методологија, заједно са рачунарском лингвистиком, представља незаобилазно средство у модерној анализи разноврсних текстова у оквиру лингвистике (Church 1991: 74). Дакле, у раду ћемо користити контрастивну лингвистику, прецизније контрастивну анализу, која се дефинише као методолошки приступ који помаже лингвисти-аналитичару да утврди у којим аспектима су испитивани језици слични, а у којима се разликују. Такође, поред знања из контрастивне лингвистике, користићемо и знања других лингвистичких дисциплина које су се нашле у контрастивном теоријско-методолошком опусу. Ту, прије свега, мислимо на когнитивну лингвистику, прагматику, лингвистику корпуса итд. Путем контрастивне лингвистике у стању смо да дубље анализирамо ова два језика што може бити од користи будућим преводиоцима и методичарима наставе. Другу, корпусну методологију, користили смо јер нам иста дозвољава квантитативно и квалитативно анализирање електронских корпуса. Ова методологија подразумијева поштовање два основна критеријума: поновљивост и доступност корпуса. Корпус треба да буде електронски скуп организованих текстова који су прикупљени на систематичан начин (McEnery 2001: 125). На овај начин корпус постаје доступан свима и свако може да провјери резултате које аутор износи.

Методологија на ужем плану

Наш корпус се састоји од великог броја текстова који су укључени у документ под називом Упитник Европске Комисије. У оквиру Упитника, највећи број докумената потпада под три регистра: законодавно-правни, друштвено-политички, дипломатски. Овакву врсту корпуса смо одабрали јер се надамо да укључује неке од најбитнијих критеријума које квалитетан корпус мора да посједује. У првом плану корпус је референтан јер представља пресјек цјелокупног модерног језика, са још једном битном карактеристиком, а то је да је сачињен од различитих стручних тематских цјелина које кроз језик представљају свако модерно друштво (економија, политика, пољопривреда, здравство, наука, спорт, образовање итд.). Друга битна карактеристика ове групе текстова јесте да је састављен радом више десетина људи из различитих дјелова Црне Горе и Србије што, надамо се, доприноси лингвистичком тј. лексичком богат-

Наслеђе 27 • 2014 • 139-154

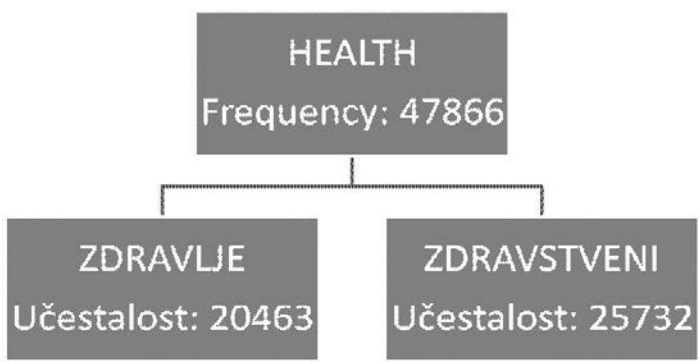
ству прикупљених текстова који ће се користити у овом раду. Одабиром оваквог корпуса избјегли смо проблем циркуларности, тј. одабир такве врсте текстова чија би анализа показала управо онакве резултате какве смо и прижељкивали. То смо учинили тако што смо одабрали преводе које су превели други колеге преводиоци, и на тај начин субјективни утицај аутора овог рада је сведен на најмању могућу мјеру. И на крају, докумената ЕУ, представљају веома битна документа за све земље бивше СФРЈ и, што је још битније за овај рад, одсликавају актуелни лингвистички тренутак и као таква заслужују одговарајућу пажњу. Наш корпус се састоји од око осам милиона ријечи са понављањем, тј. од око 65.000 ријечи без понављања. Укупно је обухваћено 128 текстова, који су у MS Word doc™ и Adobe Acrobat Reader pdf™ формату. Половина од тих докумената су на енглеском језику, док друга половина представља преводе оригиналних текстова. Обрада оволике количине информација би била веома отежана да нијесмо користили рачунарске програме који се баве оваквом обрадом текста. Главни рачунарски програми које смо користили су ТекстSTAT-2 верзија 2.8e, Heidelberg Tenka Text 0.1.3.4 и програм под једноставним називом R верзија 2.10.1. Такође смо користили WordSmith 6.0 и AntConc 3.2.4w. Након уношења ових текстова и стварања базе података, били смо у стању да анализирамо наш радни корпус. У наредном поглављу ћемо приказати неке од добијених резултата.

Табела 1. Графичко рјешење програма AntConc 3.2.4w

Резултати истраживања

Оно што се одмах дало уочити након иницијалних анализа овог корпуса јесте да је већина текстова састављена од формалних језичких структура, што смо и очекивали, с обзиром на то да овај корпус представља званични однос Црне Горе према институцијама ЕУ. Због ових елемената, можемо констатовати готово потпуно одсуство апстрактних именица или других врста ријечи јер су сви документи у оквиру корпуса фокусирани на конкретне проблематике. Ову констатацију можемо допунити и констатацијом да је већина текстова имперсонализована, што значи да су личне замјенице попут ЈА или *I* (енгл.) готово непостојеће. Такође смо открили да ријечи за које интуитивно сматрамо да би требало да су најчешће у појединим законима или поглављима Упитника то нијесу.

Тако у поглављу Упитника ЕК о Јавном здрављу и заштити потрошача, у дијелу корпуса на енглеском језику најучесталија ријеч је *HEALTH*, што се може сматрати и очекиваним с обзиром на тематику. Са друге стране у дијелу корпуса на српском језику најучесталија ријеч није ЗДРАВЉЕ, већ је то ријеч ПРОИЗВОД/А/И. Објашњење за ову разлику можемо пронаћи у стручној терминологији која је карактеристична за поље здравства. Наиме, фраза као што је *HEALTH-CARE FACILITY* је у већини случајева била преведена као ЗДРАВСТВЕНА УСТАНОВА. Са друге стране термин као што је *HEALTH-RELATED ISSUES* обично је био преведен као ПРОБЛЕМИ ВЕЗАНИ ЗА ЗДРАВЉЕ. Из ова два примјера видимо да су преводни еквиваленти за енглеску ријеч *HEALTH* подијељени на двије групе: ЗДРАВЉЕ и ЗДРАВСТВО. Енглески термин *PRODUCT* је готово увијек имао за свој преводни еквивалент нашу ријеч ПРОИЗВОД.



Наслеђе 27 • 2014 • 139-154

Коментар

Дакле, можемо закључити да најучесталија ријеч у једном дијелу корпуса не доводи закономјерно до тога да је и њен преводни еквивалент у другом дијелу корпуса најучесталији. Варијације се могу јавити услијед специфичног контекстуалног окружења, тј. преводни еквивалент зависи не само од основне ријечи већ и од контекста у којем се налази, што смо показали и кроз корпус. Такође, ово значи да интуитивни приступ изворног говорника неког језика не мора бити тачан и да, као што смо приказали у овом случају, интуиција и конкретни подаци не иду увијек руку под руку.

Имперсонализовани регистар

Оваква врста регистра, која се обично подводи под административни стил, показује велики степен шематизованости, тј. користе се разноврсне утврђене фразе како би се у што краћем времену и што јасније пренијеле идеје које неки документ носи. Најчешће се овакви административни текстови користе као вид комуникације између држава или институција. На граматичком плану овај стил је номинални, што значи да у њему доминирају именице, а процес номинализације је веома чест, тако да се тиме број глагола додатно смањује. Постојећи глаголи су обично у форми садашњег или будућег времена, у трећем лицу или безличним, пасивним конструкцијама. Чести су и декомпоновани предикати као што су:

- Донијети одлуку (умјесто одлучити),

Примјери:

учеснику - правном лицу када се ***донесе одлука*** о стечају или ликвидацији Дакле, чим се
ју или ликвидацији Дакле, чим се ***донесе одлука*** о стечају и/или ликвидацији, испуњени
.2008. године, предвидјено да се ***донесе одлука*** о избору модела за елиминисање баријер
ИИ кварталу 2008. године, биће ***донешена одлука*** о именовању Стратешког координатора
ИИ кварталу 2008. године биће ***донешена одлука*** о именовању Координатора за преког
степену у поступку у којем је ***донешена одлука*** којом је повриједјено људско право и о
раничења. У том циљу може се ***донијети одлука*** о промјени износа изражених у еурима.
да укључе грађане у процесс ***доношења одлука*** и у организацију јавних расправа. Понек

грађана у локалним процесима ***доношења одлука*** је веома ниско у Црној Гори. Међутим,

У текстовима на енглеском језику имамо другачију ситуацију. Наиме, инфинитив глагола *decide* је чешћи од одговарајућих најчешћих колокација *make decision* и *reach decision*. У овим случајевима, глаголу *decide* најчешће претходе модални глаголи и то:

V_{mod} + *bare Inf.*

- Shall decide – 91 понављање
- May decide – 90 понављања
- Can decide – 51 понављање
- Will decide – 30 понављања

➢ Извршити преглед (умјесто прегледати),

За разлику од претходне фразе нијесмо нашли нити један примјер ИЗВРШИТИ ПРЕГЛЕД у нашем корпусу, већ нас је анализа корпуса усмјерила у другом правцу. Ријеч ПРЕГЛЕД се најчешће налазила у документима који за своју тематику имају јавно здравље и заштиту потрошача (нпр. у Закону о заштити становништва од инфективних болести и у Закону о заштити потрошача). У складу са стручном терминологијом уз ријеч ПРЕГЛЕД највише линија конкорданце било је везано за:

је извршила укупно 19.631 инспекцијских ПРЕГЛЕДа, од чега 97,66% по службеној дужно
једити праћење резултата инспекцијских ПРЕГЛЕДа по одређеним групама прописа, укљ
инспектора за спровођење инспекцијских ПРЕГЛЕДа и контрола у складу са иновираном
ја евиденције о извршеним инспекцијским ПРЕГЛЕДима које врше републички инспектор
има, садржину записника о инспекцијском ПРЕГЛЕДу и друго. Закон о општем управном

За боље заштите потрошача и дефинисања правила инспекцијског надзора субјеката

ругих поремећаја здравља, 2) љекарске ПРЕГЛЕДе и друге врсте медицинске помоћи, 3) л
има претходних и периодичних љекарских ПРЕГЛЕДа запослених на радним мјестима са
Конвенција бр. 113 о љекарском ПРЕГЛЕДу морнара, усвојена 1959. године;

ела спроводе се одговарајући медицински ПРЕГЛЕДи и захвати ради процјене и смањења
их прегледа ризичне дјеце, превентивних ПРЕГЛЕДа дјеце са сметњама у развоју као и

За поље јавног здравља

У текстовима на енглеском језику имамо више варијација на тему, али три најчешће су:

and activities envisaged by the law 3) inspection CONTROL by forest inspection, environme
h trade: Environment Protection Agency (inspection CONTROL Department), Ministry of Interi
s who are in charge of carrying out the inspection CONTROL in the area of sea fisheries. A
le 73 of the Law on Tax Administration. inspection SUPERVISION in the meaning of this law
hiness of boats. Inspection in terms of inspection SUPERVISION is carried out by the inspe

За поље заштите потрошача и дефинисања правила инспекцијског надзора субјеката

и

sation and means of protection; medical EXAMINATIONs of employees which are mandatory
pply of drinking water include: medical EXAMINATIONs with a special emphasis on
cal examinations and periodical medical EXAMINATIONs every 12 months. Within the safety
nt mortality. In that sense, preventive EXAMINATIONs are conducted, which include taking of
st grade of secondary school preventive EXAMINATION of students in the third grade of secon

За поље јавног здравља

Дакле видимо да су колокацијски елементи у српском и енглеском језику различити у зависности од тематике тј. контекста у оквиру којег се ти елементи налазе. Остали примјери декомпонованих предиката, који су карактеристични за административни стил, су:

- **Поднијети извјештај (умјесто извијестити),**
- **Изразити протест (умјесто протестовати),**
- **Упутити честитке (умјесто честитати).**

Административни стил се обично дијели на пет подстилова:

1. **Законодавно-правни** , чији су жанрови закони, статути, устави, одлуке, наредбе, рјешења;
2. **Друштвено-политички** - резолуције повеље, декларације, програми, реферати, саопштења, изјаве;
3. **Дипломатски** - преписке, ноте, демарши, протоколи, меморандуми;
4. **Пословни** – уговори, дописи, фактуре, сертификати, наруџбенице, уплатнице, рачуни, спецификације;
5. **Лични** – молбе, жалбе, аутобиографије, пуномоћи, лична документа, упитници, анкете, формулари, итд.

Наш корпус је, као што смо то већ поменули, састављен од, прије свега, прве три поменуте категорије. На основу поменутих карактеристика, можемо без устезања закључити да и наш корпус одговара административном стилу. Неопходно је истаћи да ни један текст не припада чистом стилу, већ је мјешавина различитих регистара, од којих један може имати превагу. Једна од карактеристика административног стила јесте и имперсонализација, тј. увијек говоримо о државама, институцијама, министарствима, дакле, увијек о ЊИМА. Овакви документи који су по правилу веома формални, дају занимљиве резултате ако их тестирамо на личне замјенице, и контрастирамо између два језика. Тако смо нашли да се лична замјеница ОНА (никако или веома ријетко у зависности од текста) не односи на именовање припадница женског рода (+ жива), већ је готово искључиво повезана другим именицама женског рода (- жива). У енглеском језику за разлику од нашег доминира лична замјеница *IT*. Приказаћемо добијене примјере:

- поднесе Скупштини на потврђивање чим ОНА буде у могућности да се састане
- амбалажа укључена заједно с производом, ОНА ће бити укључена и за потребе
- активних мјера и/или законских поступака, ОНА ће о томе одмах обавијестити тијело

Личну замјеницу ОН нашли смо још мање пута из разлога што су именице које се чешће замјењују женског рода (држава, скупштина, влада, служба), што личну замјеницу мушког рода ОН ставља по страни:

- нуди заиста много корисних могућности. ОН комбинује двије технологије, сегментацију
- то је оно што већина људи зове Традос. ОН опслужује преводилачку меморију, омогућава
- примјену и спровођење овог Споразума. ОН ће се састајати на одговарајућем нивоу

Прва два пута говоримо о рачунарском програму, а у трећем примјеру очигледно о споразуму.

Као што смо већ рекли у енглеском језику доминирала је замјеница *IT*, а не директни преводни еквиваленти за ОН и ОНА – *HE* и *SHE*.

Наслеђе 27 • 2014 • 139-154

Зашто је то тако, одговор треба потражити у томе што је у енглеском језику другачија перцепција рода и правила налажу да се ријечи које немају природан род дефинишу са *IT*.[3] Побројаћемо само неке примјере:

- d liability companies. In that purpose, IT is necessary to initiate amendments of
- ect to separate annual programmes. When IT comes to motivation, special attention
- he company has a positive attitude when IT comes to further tertiary education of
- be a simply administrative service as IT was the case with previous personnel de

Изражавање забране

Приликом писања званичних државних докумената веома битан дио јесу права и обавезе субјеката које се дефинишу тим документима. Права и обавезе подразумијевају дозвољене и недозвољене радње које се пак различитим језичким средствима одобравају или забрањују. Посебно занимљиве, у овом контексту, јесу забране и на основу наше анализе нашли смо да и у текстовима на српском и на енглеском језику чешће се користе директни глаголи који означавају забрану (забранити или *prohibit*, *ban, proscribe*), од конструкција *NOT* + глагол допуштања (не + дозволити, не + допустити, *not allow, not approve*). Могућа су два објашњења за ову појаву:

- Директни глаголи који означавају забрану представљају дио добро установљеног шаблона,
- Психолошки ефекат (+ забрањено) је јачи од (- дозвољено).

Навешћемо неке од примјера:

- gation Magistrate, who has the right to BAN all correspondence if it harms the inve
- с *duties* постепено смањење царина PROHIBIT забранити *prohibition* забрана п
- ntains a set of laws whose stipulations FORBID discrimination, promote equality and es

Примјери превода на нашем језику:

- Магистрат, који има право да ЗАБРАНИ читаву преписку ако она штети
- с *duties* постепено смањење царина прохибит ЗАБРАНИТИ *prohibition* забрана п
- поштује низ закона који ЗАБРАЊУЈУ дискриминацију и промовишу једнакост

3 Осим у случајевима, нпр. емоционалне привржености, што у административном стилу свакако није случај.

Табела 2. Однос броја забрана[4]

Са друге стране, глаголи који су неутрални у погледу забране (поступати, бирати, *consider, use* ...), тј. они глаголи чије семантичко поље не укључује и значење забране, да би изразили забрану захтијевају негацију NOT/НЕ (или НИЈЕ/НЕЋЕ).

Коментар

Можемо закључити да је административни регистар изразито имперсонализован, али то је и очекивано. Наиме, овај регистар има један приоритет, а то је да циљној публици пренесе јасну поруку. Имперсонализација се, поред горе наведеног, огледа и у томе што иза текста увијек стоји неко министарство или држава, тако да појединац или појединци који су радили на тим преводима морају да се „сакрију" коришћењем одговарајуће лексике која то омогућава. Поред тога, административни стил обилује лексичким елементима којима се дозвољава или забрањује нешто па је то један од главних разлога зашто су глаголске или глаголско-именичке конструкције веома честе.

Нарах legomena и њихова веза са грешкама

Иза овог занимљивог латинизма крију се ријечи које су у неком тексту или групи текстова појављују само једанпут. Назив потиче од грчких ријечи (*hapax* и *legomenon* (мн. *legomena*) – речено само једанпут). У нашем корпусу од око 65.000 ријечи без понављања, 48,75 % ријечи се понављају само једном и те ријечи су веома често најуже повезане са неком струком (нпр. латинизми и врсте болести у медицини), а такође неке од ових ријечи представљају погрешно написане облике других ријечи које се иначе много чешће налазе у корпусу.

Вјероватно противној свачијој интуицији сљедећи облици ријечи се такође помињу само једном у цијелом милионском корпусу, нпр:

4 Осим у случајевима, нпр. емоционалне привржености, што у административном стилу свакако није случај.

зе, банке и друга правна лица. Поједини ***акционар*** Централне Депозитарне Агенције може има	tte Competence: Central Register of the ***Commerical*** Court 2. Making seals and stamps Timesc
раду одговарајућег документа (брошуре, ***часописи***, публикације, билтени ...) Институције	e use complex analytical techniques and ***compute*** application software. Environmental Pro
Плц (»Завод за грађевинске материјале, ***геотехничке*** и хемијске анализе« АД) – Никшић, ис ан	for himself/herself, marital partner or ***illegitimate*** partner, and for children in case that
м се дефинишу: циљеви, трајање, врсте ***гимназија***, матура, хоризонтална и вертикална прох	y use for non-agricultural purposes, is ***recultivated*** technically, chemically and biologicall
о је и кроз изјаву Центра за подводно и ***хуманитарно*** разминирање да је Црна Гора једина зем	s they should realistically decline and ***stabilise*** at 1.5% - 1.8% of Montenegrin GDP. III.

Надовезујући се на претходно поглавље пружићемо и примјере грешака које се лако могу открити у текстовима укљученим у наш корпус:

путовања могу да обављају туристичке ***аганције*** које, поред одобрења за рад (чл.11 ст.	plementation and evaluation of measures ***aggenst*** TC should be included in programs for M
у (Сл. лист ЦГ, бр.66/08), одговорна је ***Агенциоја*** за цивилно ваздухопловство Црне Горе. З	ional Protocol (OG of MN, international ***aggreements***, 2/09); Law on Veteran and Disability P
области здравства, ветерине, фармације, ***ахитектуре***, наставно-педагошког рада итд. Свако ми	d lays down requirements for ***qualtiy*** of ***agricutural*** products. Veterinary Administration is
равља у Црној Гори су отворени следећи ***центари*** и јединице за ментално здравље у зајед	alcohol per 100 kilograms of goods; 4) ***chemicalz*** and cosmetics manufacturing. Use of min
циљева других сегмената друштва. Такве ***стратегије*** су Стратегија одрживог развоја, Стратег	ce stipulates that the doctor – foreign ***citizent***, who intends to perform ***healt*** activity

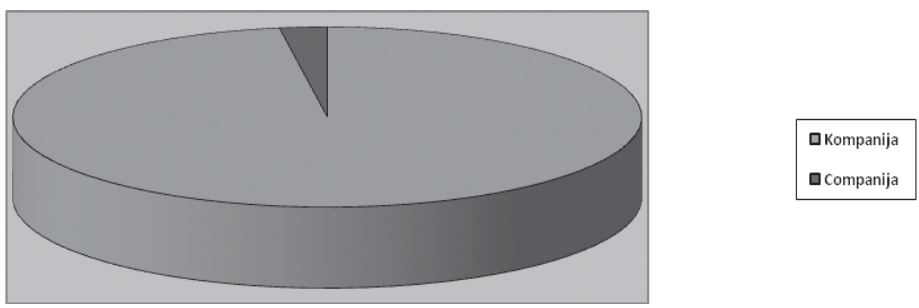
Као што се може видјети грешке у оба језика функционишу у основи на идентичне начине. У српском језику су то обично вишак или мањак знакова у ријечима, као што је то случај и у енглеском језику, оно што је познато под именом *SPELLING*.

Уочене грешке

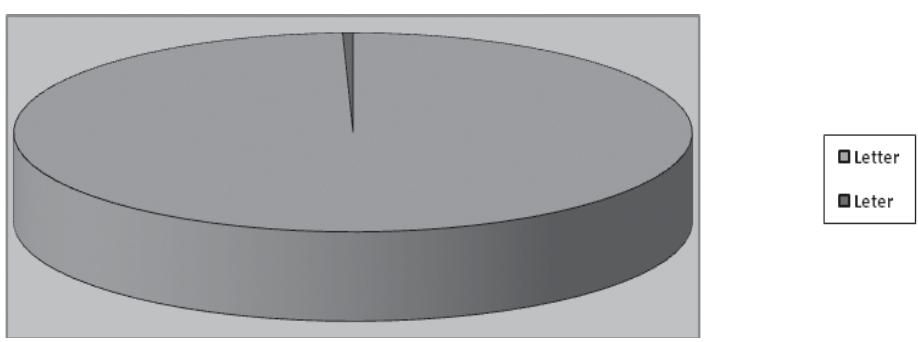
Оно што је наравно било неизбјежно јесте и то да смо наилазили на грешке у нашем корпусу. Под грешкама, прије свега, подразумијевамо типографске грешке које се могу сврстати у седам категорија:

- **Companija – примјер негативне интерференције са енглеским језиком**
- **Leter – примјер негативне интерференције са српским језиком**
- **Ћлан – грешке у избору дијакритичког знака**
- **СкупштинаЦрне Горе – грешке у прављењу размака између ријечи**
- **звјештаји – недостатак слова – непотпуна ријеч**
- **стартегије – неправилан редослијед слова**
- **репубблика – непотребно дуплирање слова**

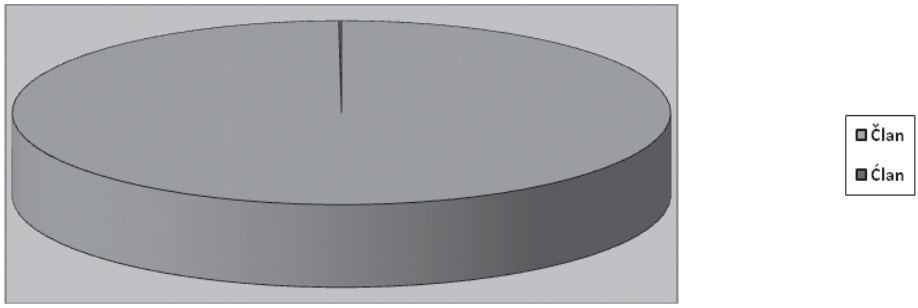
Оно што смо такође пронашли јесте да се све овакве грешке понашају по истом принципу, а то је да се налазе ван зоне учесталости понављања која је карактеристична за исправне верзије датих ријечи. Илустроваћемо неке примјере дијаграмима на којима ће се јасно видјети шаблон понашања типографских грешака у нашем корпусу.



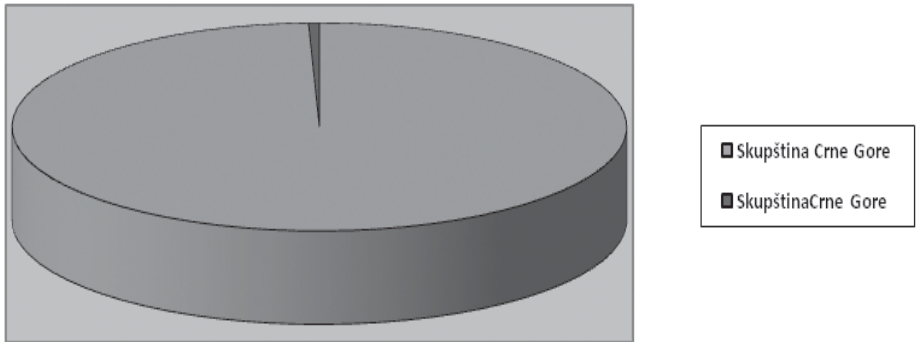
Дијаграм 1. - Однос учешћа грешке бр. 1 – негативне интерференције енглеског језика на домаћу ријеч



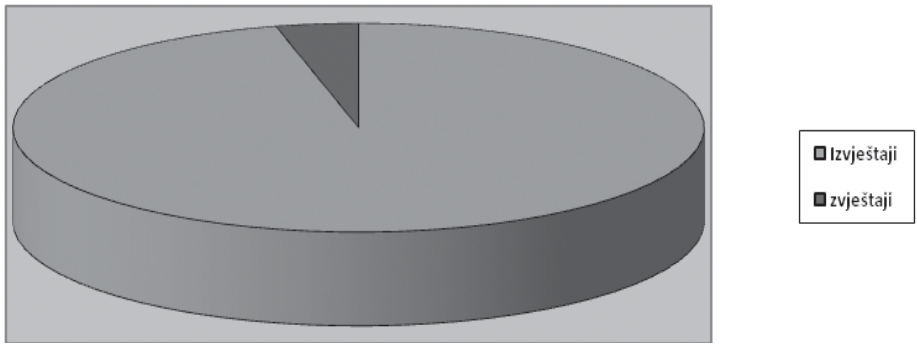
Дијаграм 2. – Однос учешћа грешке бр. 2 – негативне интерференције српског језика на енглеску ријеч



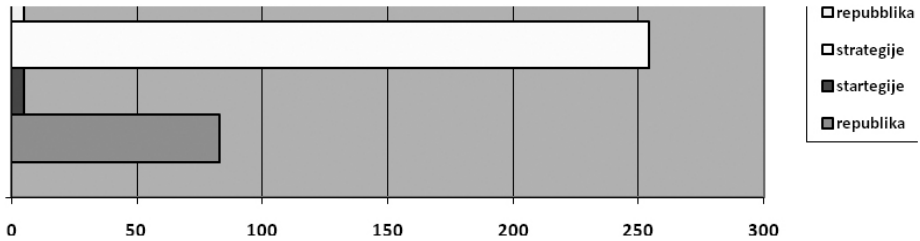
Дијаграм 3. – Однос грешке бр. 3 – погрешан избор дијакритичког знака



Дијаграм 4. – Однос грешке бр. 4 – неодвајање ријечи



Дијаграм 5. – Однос грешке бр. 5 – непотпуна ријеч



Дијаграм 6. – Однос грешке бр. 6 и 7 – неправилан ред слова у оквиру ријечи и непотребно дуплирање знакова

Јасно се може уочити шаблон где се погрешна ријеч јавља вишеструко мање пута у односу на исправни облик те ријечи, што је и очекивано. Ми смо овдје издвојили само неке од примјера грешака које се могу наћи, али оно што је занимљиво јесте да и све остале грешке, без обзира на своје карактеристике, прате идентичан начин понашања, као што смо показали на Дијаграмима 1 – 6. Све грешке ће се увијек налазити ван оптималног броја понављања и то је правилност и универзална појава како за наш тако и за енглески језик.

Коментар

Ријечи које се понављају само једанпут о оквиру милионских корпуса могу се посматрати на два начина: или представљају „егзотичне" ријечи које већина говорника неког језика никада и не чује и не види и не користи, или су те ријечи у директној вези са ортографским грешкама. Ова друга категорија је, условно речено, значајнија. Једна од области која интензивно користи овакве резултате јесте индустрија рачунарских програма. Рачунарски програми (енгл. *SPELL CHECK*) препознају овакве грешке на основу статистичке анализе својих база података.

Закључак

Као што се може закључити једна од основних премиса овог рада јесте жеља да се у контрастивну анализу унесу елементи корпусне и рачунарске лингвистике. На тај начин смо спојили провјерени квалитет контрастивног метода са модерним карактеристикама корпусне и рачунарске анализе. Оне су, прије свега, донијеле квалитативни и квантитативни помак у односу на истраживања која су се базирала искључиво на контрастивном приступу. Кроз многобројне примјере, надамо се, показали смо које су то основне карактеристике административног стила писања који је доминантан стил када се ради о међудржавним односима. Корпусна и рачунарска лингвистика, све више и више, добијају на свом

значају, али је неопходно напоменути да и оне као такве имају својих мана, а једна од најочигледнијих јесте да нам могу пружити резултате, али нам не могу рећи какав је квалитет тих резултата у смислу њихове, нпр. истинитости, али, и поред тога, приказали смо да корпус може бити веома корисно средство у лингвистичким анализама. Анализирали смо разлике у административном дискурсу текстова на српском и енглеском језику које су биле условљене не само формалним граматичко-структуралним разликама између ова два језика, већ и културолошким и системско-правним разликама које се пресликавају и утичу на њих.

Поред тога, можемо закључити да је корпус ипак поузданије оруђе у лингвистичким анализама него што је то интуиција изворног говорника, што смо показали на примјерима у вези са ријечима *здравство* и *здравствени*, у српском језику и *health* у енглеском језику, а такође смо видјели да се неке ријечи понављају свега пар пута или се уопште не налазе у корпусу иако вјероватно већина изворних говорника не би тако оцијенила на основу своје интуиције.

Кроз контрастирање лексичких и реченичних елемената из српског језика и њима одговарајућих преводилачких еквивалената у енглеском језику анализирали смо врсту дискурса који се користи у оваквим типовима административних докумената (дискурсна обиљежја).

Литература

Agresti 2002: A. Agresti, Categorical Data Analysis. Hoboken, NJ: Wiley.

Aston (ed.) 2001: G. Aston, Learning With Corpora. Bologna. CLUEB.

Baayen 2008: R. Baayen, Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX Lexical Database (Release 2). Philadelphia, PA: Linguistic Data Consortium.

Barnbrook 1996: G. Barnbrook, Language and Computers. Edinburgh: Edinburgh University Press.

Biber 1988: D. Biber, Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber 1998: D. Biber, Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Bod 2003: R. Bod, Probabilistic Linguistics. Cambridge, MA: MIT Press.

Bugarski 1972: R. Bugarski, Jezik i lingvistika, Nolit, Beograd.

Bugarski 1986: R. Bugarski, Terminologija kontrastivne lingvistike, u: Kontrastivna jezička istraživanja, III simpozijum (Novi Sad, 6. i 7. decembar 1985), *Zbornik radova*, Univerzitet u Novom Sadu, Filozofski fakultet, Novi Sad, 383–390.

Bugarski 1990: R. Bugarski, Integralna kontrastivna analiza, u: Kontrastivna jezička istraživanja, IV simpozijum (Novi Sad, 8. i 9. decembar 1989), *Zbornik radova*, Filozofski fakultet, Novi Sad, 58–62.

Church 1991: K. Church, Using statistics in lexical analysis. In Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, ed. Uri Zernik, 115–164. Hillsdale, NJ: Lawrence Erlbaum.

McEnery 2001: T. McEnery, *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Sinclair 1991: J. Sinclair, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Igor D. Ivanović

CORPUS AND COMPUTATIONAL LINGUISTICS ON THE EXAMPLE OF THE EC QUESTIONNAIRE

Summary

This paper deals with the analysis of one part of the European Commission Questionnaire. From the complete analysis in our doctoral thesis, two aspects have been chosen for the purposes of this paper. The first one is the analysis of impersonalised register, while the second aims to find the connexion between *hapax legomena* and the mistakes we have spotted in the abovementioned corpus. We will show how the use of corpora and computers can shed new light onto traditional linguistic research. Furthermore, this paper should also be understood as a way of promoting the Corpus and Computational linguistics which are yet to find their full application within linguistic research.

Key words: corpus, linguistic analysis, computers, the EC Questionnaire

Примљен у јуну 2013.
Исправљен 08.02.2014.
Прихваћен 22.02.2014.