AKTET **ISSN 2073-2244**
Journal of Institute Alb-Shkenca **www.alb-shkenca.org**
Revistë Shkencore e Institutit Alb-Shkenca

BROWN CORPUS OF ALBANIAN LANGUAGE AS THE BASIS FOR EXTRACTING KEY-WORDS OF THE DIFFERENT TEXTS IN ALBANIAN LANGUAGE
KORPUSI NJËMILIONFJALËSH I GJUHËS SHQIPE SI BAZË PËR NXJERRJEN E FJALËVE ÇELËS TË TEKSTEVE TË NDRYSHME NË GJUHËN SHQIPE

NEBI CAKA[1], ALI CAKA[2]
[1] Fakulteti i Inxhinierisë Elektrike dhe Kompjuterike
[2] Fakulteti i Filologjisë, Universiteti i Prishtinës, Prishtinë, KOSOVË
nebi.caka@uni-pr.edu

AKTET V, 4: 714-719, 2012

PERMBLEDHJE
Korpusi njëmilionfjalësh i gjuhës shqipe i përpiluar nga N. Caka dhe A. Caka më 2006, sipas korpusit përkatës të anglishtes, të njohur si 'Korpusi i Braunit', është një bazë e mirë për hulumtime të ndryshme gjuhësore në fushë të leksikut, por jo vetëm të tij. Atë e shfrytëzoi, për herë të parë, vetë njëri nga autorët (A. Caka) gjatë punimit të tezës së tij të magjistraturës shkencore, me titull "Kontribut për fjalorin themelor të gjuhës shqipe", të cilën e mbrojti në Fakultetin e Filologjisë të Universitetit të Prishtinës më 2007. Duke marrë për bazë këtë korpus, e duke përdorur programin kompjuterik përkatës, të njohur si konkordancier, për përcaktimin e dendurisë së përdorimit të fjalëve, do të bëhet përqasja e teksteve të ndryshme, a korpuseve të tjera më të vogla, me këtë korpus me qëllim të nxjerrjes së fjalëve çelës të këtyre teksteve.
Fjalët kyçe: korpusi njëmilionfjalësh, konkordancieri, fjalët-çelës, denduria-e-fjalëve, faktori-i-çelësshmërisë.

SUMMARY
Onemillion-words corpus of the Albanian language, compiled by N. Caka and A. Caka in 2006, according to the relevant corpus of English, known as 'The Brown Corpus', is a good basis for various linguistic research. It is used, for the first time, by one of the authors (A. Caka) during his magisterial thesis research, entitled "Contribution to the basic vocabulary of the Albanian language", which he defended at the Faculty of Philology of the University of Prishtina in 2007. Using the corresponding computer program, known as concordancier, to determine the density of using words, the 'Brown' Corpus of Albanian language is used for extracting key-words of the different texts in Albanian language.
Key words: Brown Corpus, concordancier, key-words, word-frequency, keynees-factor

1. Hyrje

Zhvillimi i përshpejtuar i teknologjisë së informacionit ka mundësuar zhvillimin edhe të teknologjive gjuhësore të cilat sot ofrojnë rrugë të reja për studime gjuhësore. Analizat gjuhësore kanë marrë një kah tjetër falë korpuseve tekstore të cilat gjerësisht janë duke u përdorur për të llogaritur faktorin e çelësshmërisë (*keynees-factor*) dhe përzgjedhjen e fjalëve çelës të teksteve a përmbajtjeve të caktuara. Faktori i çelësshmërisë mundëson gjetjen e shpejtë të përmbajtjes së caktuar në internet gjatë kërkimit nëpërmjet fjalëve çelës të caktuara.

2. Korpusi (origjinal) i Braunit dhe Korpusi 'i Braunit' i gjuhës shqipe

Korpusi i Braunit i anglishtes përmban 1 milion fjalë të marra nga 500 tekste me nga 2000 fjalë me një gamë të gjerë të temave. Ai ka 31448

trajta fjalësh të ndryshme. Në vendin e parë është ***the***, e pasojnë ***of, and, to, a***.
Gjatë përpilimit të korpusit njëmilionfjalësh të gjuhës shqipe kemi zbatuar parimet e përpilimit të Korpusit të Braunit të anglishtes.
Në Korpusin tonë 'të Braunit' të shqipes, në vendin e parë është ***të,*** e pasojnë ***e, në, dhe, i***.

3. Fjalët çelës dhe nxjerrja e tyre

Në një punim tonin të mëhershëm kemi thënë se parapëlqejmë termin "fjalët çelës" kundrejt termit "fjalët kyçe", jo vetëm pse fjala 'kyç' është me origjinë sllave (nga: ključ = mbyllës 'çelës'), por se fjala "kyç" më parë se 'çelës' e ka kuptimin e 'drynit', prandaj 'çelës' më mirë se 'kyç' i shkon këtij togfjalëshi në rastin kur me të dëshirojmë të shënjojmë fjalët që janë çelës (angl. *key*, fr. *clef*, it. *chiave* etj.), e jo *dry* (mbyllës) i një teksti/shkrimi të një autori [Caka & Caka, 2007].
Koncepti i termit "fjalët çelës" (angl. *key-words*) daton nga viti 1976, kur Raymond Williams, teoricien letrar dhe kritik kulturor britanik, boton në Londër dhe në Nju Jork një vëllim të vogël me titullin *Keywords: A Vocabulary of Culture and Society* [Williams, 1976], që në fakt ishte shtojca që i kishin thënë ta hiqte nga libri i tij *Culture and Society: 1780-1950*, të botuar në Nju Jork më 1958.
Derisa më 1958, me fjalë çelës nënkuptonin fjalët që shërbenin për të zbuluar strukturën e brendshme të gjykimit (rezonimit) të një autori, sot me fjalët çelës rëndom nënkuptojmë fjalët që na shërbejnë për të kërkuar në Internet dokumente, ngjarje a faqe të ndryshme. Me avancimin e progresit të epokës së Internetit, kuptimi i dytë është duke mbizotëruar gjithnjë e më shumë [Lissack, 2010].
Në artikujt shkencorë, fjalët çelës janë fjalë a togfjalësha që paraqesin koncepte të caktuara (ndaj rëndom janë emra ose fraza emërore) të cilat më së miri e përshkruajnë përmbajtjen e artikullit për qëllim të indeksimit të tij dhe të kërkimit në Internet. Ato radhiten sipas renditjes alfabetike dhe nuk mund të jenë më shumë se dhjetë, rëndom katër deri në shtatë. Jepen menjëherë pas përmbledhjes/abstraktit.

Në një analizë të Raporteve vjetore të bërë nga Klímová (2004) në një grup prej 50 fjalësh çelës shumica janë të formuara nga emrat (35, a 70 %), të ndjekura nga mbiemrat (8, a 16 %), shkurtesat (6, a 12 %) dhe numërorët (1, a 2%) dhe asnjë folje.
Në gjuhësinë e korpusit, me fjalë çelës nënkuptojmë fjalët të cilat shfaqen me një denduri (shpeshti) jo të zakonshme (shumë të madhe a shumë të vogël) në një tekst a në një korpus tekstor (grup tekstesh), kur lista e fjalëve të tekstit që shqyrtohet krahasohet me listën e fjalëve të një korpusi referent, i cili na shërben si standard krahasimi a referimi. [Bednarek 2007] Kështu, p.sh., gjatë nxjerrjes së fjalëve çelës të një romani a të një drame a poezie të një autori, si korpus referimi mund të shërbejë lista e fjalëve a fjalësi i tërë veprës letrare të atij autori, por edhe fjalësi vetëm i romaneve (dramave, poezive), por dhe korpusi 'i Braunit', a ndonjë korpus tjetër i caktuar edhe më i madh, por jo dhe tepër i madh [Scott & Tribble 2006].
Për nxjerrjen e fjalëve çelës të një korpusi të caktuar, duke krahasuar dendurinë e paraqitjeve të fjalëve në këtë korpus me dendurinë e paraqitjeve të fjalëve në korpusin origjinal njëmilionfjalësh të Braunit të anglishtes janë përpiluar edhe programe kompjuterike të caktuara, disa prej të cilave jepen dhe falas, si p.sh. 'KeyWords Extractor v. 1', i cili mundëson nxjerrjen e fjalëve çelës nga korpuse të vogla, deri në 50 mijë fjalë, por ka mundësinë që ta bëjë këtë edhe nga korpuse prej 1,5 milionë fjalëve. Për testimin e këtij programi është përdorur një tekst (punim shkencor i botuar) në anglisht me rreth 3000 fjalë (Nebi Caka, Astrit Hulaj, *The analysis of different FTTH architectures and possibilities of their implementation in Kosova*), e për testimin e korpusit tonë 'të Braunit' është përdorur i njëjti tekst i përkthyer në shqip (Nebi Caka, Astrit Hulaj, *Analiza e arkitekturave të ndryshme FTTH dhe mundësitë e zbatimit të tyre në Kosovë*).

1) 18324.00 kosovo	(28) 567.09 architecture	(55) 86.61 representation
(2) 15205.00 ftth	(29) 451.42 optical	(56) 72.53 loss
(3) 4678.00 optic	(30) 390.00 microwave	(57) 70.91 clients
(4) 3509.00 gpon	(31) 377.29 network	(58) 68.82 neighborhoods
(5) 2729.00 technologies	(32) 375.41 fiber	(59) 68.82 cabinet
(6) 2729.00 telecommunications	(33) 316.75 transmission	(60) 65.00 laying
(7) 1949.00 graphical	(34) 292.50 implement	(61) 65.00 installation
(8) 1949.00 providers	(35) 272.90 passive	(62) 65.00 fulfillment
(9) 1949.00 infrastructure	(36) 260.00 kilometers	(63) 64.98 measurements
(10) 1559.00 splitter	(37) 260.00 implementing	(64) 61.58 shorter
(11) 1559.00 internet	(38) 259.83 users	(65) 60.00 wires
(12) 1413.25 implementation	(39) 239.92 copper	(66) 52.29 capacity
(13) 1170.00 reflectance	(40) 226.67 technology	(67) 51.01 distance
(14) 1170.00 epon	(41) 222.71 cable	(68) 43.33 faster
(15) 1170.00 bpon	(42) 173.22 disadvantages	(69) 42.37 losses
(16) 1170.00 multimedia	(43) 167.07 advantages	(70) 41.79 customer
(17) 1170.00 architectures	(44) 162.46 access	(71) 40.80 companies
(18) 780.00 ethernet	(45) 156.00 grouped	(72) 37.74 panel
(19) 780.00 broadband	(46) 137.59 networks	(73) 37.58 speed
(20) 780.00 end-user	(47) 135.61 fibers	(74) 37.14 strategy
(21) 780.00 telephony	(48) 111.43 bending	(75) 34.40 realized
(22) 780.00 implemented	(49) 111.43 configuration	(76) 34.40 services
(23) 780.00 electromagnetic	(50) 111.36 analyzed	(77) 33.91 assessment
(24) 780.00 bandwidth	(51) 97.50 ensure	(78) 32.50 realization
(25) 780.00 videogames	(52) 97.50 villages	(79) 30.00 justify
(26) 780.00 iptv	(53) 97.50 termination	(80) 30.00 electricity
(27) 779.50 cables	(54) 90.00 patch	

Tabela 1. Lista e fjalëve çelës potenciale të tekstit në gjuhën angleze të renditura sipas faktorit të çelësshmërisë (më të madh se 30)

4. Rezultatet

Fjalët çelës të mundshme (potenciale) të teksti/punimit "The analysis of different FTTH architectures and possibilities of their implementation in Kosovo" (Nebi Caka, Astrit Hulaj) (me 2764 fjalë), të nxjerra me KeyWords Extractor v. 1, jepen më poshtë, të renditura sipas numrit virtual të paraqitjeve në tekst a 'faktorit të çelësshmërisë' ("*keyness factor*"), i cili llogaritet sipas shprehjes:

Faktori i çelësshmërisë = "*numri i paraqitjeve të fjalës në tekst* pjesëtuar me *numrin e gjithmbarshëm të fjalëve në tekst*" shumëzuar me "1.000.000 pjesëtuar me *numrin e paraqitjeve të fjalës në korpusin e Braunit* (përkatësisht në korpusin *Caka & Caka*)".

(1)	12935.32	FTTH	(38)	316.60	humbje
(2)	3980.10	PON	(39)	304.04	kapacitet
(3)	2985.07	GPON	(40)	295.82	fije
(4)	2653.40	bakër	(41)	265.34	ofroj
(5)	2653.40	transmetim	(42)	248.76	realizim
(6)	2321.72	pasiv	(43)	248.76	sinjal
(7)	2321.72	kosto	(44)	236.91	tel
(8)	1990.05	realizoj	(45)	236.91	elektrik
(9)	1990.05	ofrues	(46)	199.00	mirëmbajtje
(10)	1990.05	bazoj	(47)	165.84	telekomunikacion
(11)	1575.46	arkitekturë	(48)	132.67	paraqitje
(12)	1326.70	reflektancë	(49)	124.38	tërësishëm
(13)	1326.70	PTK	(50)	122.20	tabelë
(14)	1326.70	shpërndarës	(51)	118.46	konsumator
(15)	1326.70	disavantazh	(52)	97.55	shërbim
(16)	1326.70	kabllo	(53)	90.46	kryej
(17)	1260.36	optik	(54)	88.45	artikull
(18)	995.02	multimedial	(55)	80.68	shpejtësi
(19)	995.02	dollap	(56)	76.54	komunikim
(20)	995.02	EPON	(57)	75.38	rrit
(21)	995.02	BPON	(58)	73.71	përshtatshëm
(22)	862.35	avantazh	(59)	71.07	investim
(23)	829.19	matje	(60)	66.33	analizë
(24)	663.35	medium	(61)	66.33	shtrirje
(25)	663.35	analizoj	(62)	62.19	përdor
(26)	552.79	grafik	(63)	62.19	rural
(27)	497.51	trend	(64)	58.53	ndërtesë
(28)	497.51	ndarës	(65)	57.19	ardhme
(29)	497.51	ofrim	(66)	55.28	afatgjatë
(30)	483.69	teknologji	(67)	52.65	kompani
(31)	438.67	rrjet	(68)	52.37	internet
(32)	422.13	distancë	(69)	41.46	pozitë
(33)	379.06	përdorues	(70)	39.80	strategji
(34)	379.06	akses	(71)	38.57	aktual
(35)	331.67	infrastrukturë	(72)	37.56	Kosovë
(36)	331.67	përmbushje	(73)	36.85	krahasim
(37)	321.62	zbatim	(74)	33.17	lagje

Tabela 2. Lista e fjalëve çelës potenciale të tekstit në gjuhën shqipe të renditura sipas faktorit të çelësshmërisë (më të madh se 30)

Shënim:

1. Nuk janë marrë parasysh fjalët me më pak se dy paraqitje (199).

2. Me qëllim të nxjerrjes së fjalëve çelës të vlefshme (valide) janë injoruar (eliminuar) të ashtuquajturat fjalët jokuptimplota, si 'to', 'the'

etj., të cilat janë hequr me ndihmën e listës së fjalëve "stop words" (së këndejmi .
3. Për fjalët që dalin në tekstin që shqyrtohet e nuk janë në korpusin e Braunit është marrë se kanë nga një paraqitje.

Në rastin e tekstit në gjuhën shqipe (me gjithsej 3015 fjalë) si korpus referimi është marrë korpusi njëmilonfjalësh (Caka & Caka 2007).
Nëse i krahasojmë tabelat 1 dhe 2 vërejmë se fjala **Kosovë** (**Kosovo)** gjendet në vendin e parë në Tabelën 1, me 'numrin e paraqitje virtuale' a 'faktor të çelësshmërisë' të barabartë me: 47/2565 x 1,000,000/1 = **18324.00**, ndërsa në vendin e 72-të në Tabelën 2, me 'faktor të çelësshmërisë' të barabartë me: 47/3015 x 1,000,000/415 = **37.56** (ku 47 është numri i paraqitjeve të fjalës Kosovë në tekstin shqip që shqyrtohet, ndërsa 415 numri i paraqitjeve të fjalës Kosovë në korpusin e gjuhë shqipe 'Caka & Caka').
Sa më i madh 'faktori i çelësshmërisë' aq më e madhe është gjasa që një fjalë të jetë fjalë çelës.
Tekstet e vogla mund të japin krahasime jo të besueshme. Nga ana tjetër, siç argumentojnë Chun dhe Nation (2004), një faktor i çelësshmërisë më i vogël se 50 është jointeresant. Në rastin tonë, meqë kemi të bëjmë me një tekst jo aq të madh, faktori i çelësshmërisë është marrë të jetë 30, dhe për këtë faktor kemi 74 fjalë çelës të mundshme nga një total prej 3015 fjalësh, a 74/3015 = 0,02 = 2 %.
Në një analizë të bërë nga Liu, Wu dhe Zhou (1999), më se 60% e përdoruesve mendojnë se fjalët çelës dhe nxjerrja e përqindjes së tyre të përdorura së bashku mund të ndihmojnë për të shkruar përmbledhjen/abstraktin e një artikulli.
Fjalët çelës (pas Abstraktit anglisht, përkatësisht pas Abstraktit/Përmbledhjes shqip), të cilat i japin vetë autorët në tekstet e marra në shqyrtim janë: Access network, FTTH, PON, GPON, fiber optic, Kosovo, telecommunications, architecture; përkatësisht: Rrjeti i aksesit, FTTH, PON, GPON, fija optike, Kosova, telekomunikacioni, arkitektura.

Bibliografia

BEDNAREK, Monica (2007). Teaching English Literature and Linguistics Using Corpus Stylistic Methods, *Bridging Discourses: ASFLA 2007 Online Proceedings*.

CAKA, Ali (2007). *Kontribut për fjalorin themelor të gjuhës shqipe* punim magjistrature; mentor: Rexhep Ismajli), Prishtinë: Universiteti i Prishtinës, Fakulteti i Filologjisë.

CAKA, Nebi (1985). *Analizë statistike dhe stilistike e gjuhës së Lasgush Poradecit* (punim magjistrature; mentor: Rexhep Ismajli), Prishtinë: Universiteti i Prishtinës, Fakulteti Filozofik.

CAKA, Nebi dhe CAKA, Ali (2006), *Korpusi njëmilionfjalësh i gjuhës shqipe*, Prishtinë.

CAKA, Nebi dhe CAKA, Ali (2007), "Konkordancat dhe fjalët-çelës të poezisë së Ndre Mjedës*"*, *AKTET e Takimit Ndërkombëtar Vjetor të IASH*, Tiranë – Prishtinë – Shkup: Instituti Alb-Shkenca, Vëll. I, Nr. 1, f . 129-138.

CHUNG, T. and Nation, I.S.P. (2004) Identifying technical vocabulary. *System* 32, 2 251-263.

HANNERZ, Ulf (1997), Flows, boundaries and hybrids: keywords in transnational anthropology, Published in Portuguese as "Fluxos, fronteiras, híbridos: palavras-chave da antropologia transnacional", *Mana* (Rio de Janeiro), 3(1): 7-39.

KLÍMOVÁ, Blanka (2004). Lexical Analysis of Annual Reports, *Theory and Practice in English Studies*, Proceedings from the Seventh Conference of English, American and Canadian Studies (Literature and Cultural Studies), Edited by Pavel Drábek and Jan Chovanec, Brno: Masaryk University, p. 77-86.

KUČERA*,* H. and W. FRANCIS (1967) *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

LISSACK, Michael R. (2010), *Remedy101: Don't let efficiency overwhelm resilience*, Institute for the Study of Coherence and Emergence, September 2010.

LIU, James; WU, Yan; ZHOU, Lina (1999), A Hybrid Method for Abstracting Newspaper Articles, *Journal of the American Society for Information Science*. 50(13):1234–1245.

MUDRAYA, Olga (2006). Engineering English: A lexical frequency instructional model, *English for Specific Purposes* 25 235–256.]

SCOTT, M. & TRIBBLE, C. (2006), *Textual Patterns: keyword and corpus analysis in language education*, Amsterdam: Benjamins.

SCOTT, M. (1996). WordSmith tools. Oxford: Oxford University Press. Available from http://www.lexically.net/wordsmith/.

SCOTT, M. (1997). PC analysis of key words – and key key words. System, 25(2), 233–245.

SOUBBOTIN, M.M., SOUBBOTIN, S.M. (2001). *Patterns of Potential Answer Expressions as Clues to the Right Answers*. In: 10th Text Retrieval Conference (TREC 2001), Gaithesburg, MD, pp. 293-302.

WILLIAMS, Raymond (1976). *Keywords: A Vocabulary of Culture and Society.* New York: Oxford UP.