Uporaba korpusov

Tomaž Erjavec
Korpusno jezikoslovje
UNG
2008/2009

24. 10. 2008

Pregled predavanja

1. uvod
2. konkordančniki in regularni izrazi
3. pojavnice in različnice
4. konkordance
5. besedni seznami
6. ključne besede in termini
7. kolokacije

Kaj lahko jezikoslovno analiziramo?

A. uporabo neke jezikoslovne lastnosti (leksikalne ali gramatične)
   1. jezikoslovna asociacija lastnosti (leksikalna ali gramatična)
   2. nejezikoslovna asociacija lastnosti (distribucija po registrih, dialektih, časovnih obdobjih)
B. jezikovno raznolikost (po registrih, dialektih, časovnih obdobjih)
   1. vzorci jezikoslovnih asociacij

Bolj enostavno: vse kombinacije leksikalnih, gramatičnih in nejezikoslovnih lastnosti

Primeri posameznih analiz

A. uporaba besede “zgoščenka” (leksikalna lastnost) ali uporaba glagolnikov (gramatična lastnost)
   1. s katerimi besedami se najbolj pogosto uporablja ali katere besedne vrste se pojavljajo v njeni okolici
   2. kako se uporablja v tehničnih/tehničnih besedilih ali kako se uporablja po letih
B. v čem se razlikujejo tehnična od netehničnih besedil (vrsta besedila) ali v čem se razlikujejo besedila pred 1991 od tistih po 1991 (časovno obdobje)
   1. kako se razlikuje leksika ali
      kako se razlikuje uporaba besednih vrst

A.1 “zgoščenka” in leksikalna okolica: Word Sketches


zgoščenka Fida PLUS 620m freq = 14595

a_modifier	3563	1.4
kompilacijski	40	53.47
priložen	99	51.81
nov	1226	48.64
piratski	36	40.21
multimedijski	48	40.01
glasben	154	35.73
spremljajoč	35	34.0
samostojen	61	29.54
avtorski	38	27.85
računalniški	50	26.99

post_z	1342	8.4
naslov	587	62.42
glasba	166	44.15
posnetek	61	35.25
skladba	42	33.1
pesem	61	31.08

is_obj4	2426	6.0
izdati	839	62.56
posneti	267	47.86
predstaviti	260	35.61
usesti	57	30.84
pripravljati	65	21.69
podariti	31	21.55
kupiti	57	20.01
prejeti	39	17.25
predstavljati	33	13.24
pripraviti	39	10.79

prec_sin	1077	5.5
skladba	35	31.73
izdi	51	28.9
posneti	45	27.42
pesem	33	24.55
izdati	41	22.83
glasba	31	22.61
bit	163	21.23
najti	40	16.26

prec_z	498	3.1
nagradi	106	51.51
glasba	35	27.99

gen_2	1529	2.4
predvajalnik	131	61.11
CD	143	58.3
ovitek	39	42.44
naslovnica	39	39.13
izdaja	74	37.92
izid	82	37.4
predstavitev	82	35.94
snemanje	44	33.73
promocija	34	29.79
prodaja	48	25.41

is_obj2	268	2.2
izdati	35	28.26

prec_prep	2545	1.9
na	1559	38.66
z	539	25.97
za	216	15.32

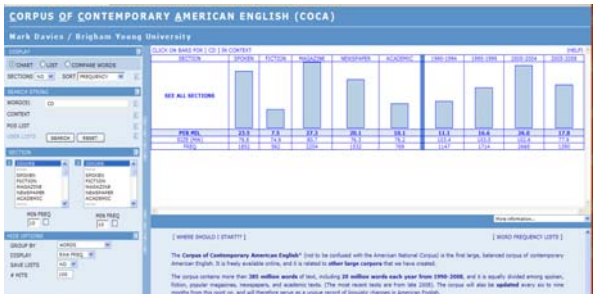
coord	2463	1.7
kaseta	1153	98.57
videokaseta	49	45.08
plošča	52	26.89
knjiga	82	26.28
drug	34	18.82
on	40	15.72

is_subj	1046	1.7
iziti	292	56.08
priznati	34	22.83
vsebovati	36	22.03
namenjti	31	16.67

gen_1	589	0.9
licenca	65	48.14
založba	75	45.0
ansambel	54	41.97
skupina	58	27.05

prec_verb	2195	0.7
izdati	655	58.89
posneti	209	44.66
iziti	130	37.23
snemati	41	27.15
predstavti	113	25.3
kupiti	45	18.12
pripravljati	38	16.65
prodati	30	12.59
pripraviti	34	10.21
dobiti	31	7.11

A.2 “CD” v COCA/BYU po registrih in časovnih obdobjih



Uporaba korpusov

Opozorilo: orodja niso popolna!

- kaj konkordančnik razume kot “besedo”
- kako so besede označene
- kakšne vrednosti vrnejo statistične formule
- rezultate, dobljene iz korpusov, je potrebno kritično ovrednotiti

Konkordančniki

- najbolj pogosto orodje za raziskovanje korpusov
- poleg samih konkordanc ponavadi ponujajo še druge funkcionalnosti
  - frekvenčni seznami, statistične obdelave
  - sortiranje, filtriranje
  - izbira podkorpusov po metapodatkih
  - hramba, izpis najdenega

Vrste konkordančnikov

- nekatere konkordančnike dobimo ali kupimo in namestimo na svoj računalnik
  - sami si moramo zagotoviti korpus(e)
- mrežni konkordančniki
  - ni potrebe po instalaciji, potrebujemo pa mrežno povezavo
  - ponujajo (enega, več) velikih korpusov
  - večina pa ne nudi možnosti za nalaganje lastnih korpusov (izjema: SketchEngine, vendar plačljiv)
- poizvedovalni jeziki in funkcionalnosti se razlikujejo od orodja do orodja

Poizvedovalni jezik

- vsak konkordančnik omogoča “mehko iskanje”, oz. iskanje preko regularnih izrazov
- kjer so korpusi označeni, je mogoče iskati tudi po oznakah (npr. lema in oblikoskladenjska oznaka)
- primer konkordančnika z bogatim iskalnim jezikom, sicer pa bolj revno funkcionalnostjo:
  - CWB-IJS: iKorpus (demonstracija)

Regularni izrazi

- regularne izraze uporablja večina konkordančnikov (imenujejo se tudi “mehko iskanje”)
- uporabljajo jih tudi urejevalniki besedil in mnogo programskih jezikov (grep, awk, Perl, Ruby,…)
- regularni izraz prepozna (mogoče neskončno) množico nizov
- sestavljeni so iz literalov in operatorjev:
  literali: npr. *a,b,c,č,d,…,z,ž*
  operatorji: konkatenacija, disjunkcija, ponavljanje, združevanje

Osnovni primeri

- konkatenacija (implicitna):
  /abc/ prepozna {*abc*}
- disjunkcija:
  /ab|bc/ prepozna {*ab, bc*}
- ponavljanje:
  - ničkrat ali enkrat:
    /ab?/ prepozna {*a, ab*}
  - ničkrat ali večkrat:
    /ab*/ prepozna {*a, ab, abb*, …}
  - enkrat ali večkrat:
    /ab+/ prepozna {*ab, abb*, abbb,…}
- združevanje:
  /(ab?)|c / prepozna {a, *ab, c*}

Razširitve sintakse

- katerikoli literal: “.”
  npr. /abc./
- pogosta uporaba: “.*”
  npr. /abc.*/
  - dosti programov “.*” okrajša na “*”
- skupine literalov: “[…]”
  npr. / [fgm]iga/ prepozna {figa, giga, miga}
- negirana množica literalov [^…]
  npr. /abc[^def]ghi/ prepozna {*abcgghi, abchghi, abcighi, …, abczghi, abcžghi*}
- ponavljanje: “{n,m}”
  npr. /a{2,5}/ prepozna {*aa, aaa, aaaa, aaaaa*}

Primeri za iKorpus

- miza, miz., miz.?, miz.*
- miz[a,e,i,o], miz(a|e|i|o|ama|ah|ami)
- .*pisati, …pisati
- .*gled.*, pod.*, .*anje
- [aeiou]+

Naloge iz regularnih izrazov

Napišite naslednje iskalne pogoje:

- besede, ki se začnejo na “miš”
- besede, ki vsebujejo “miš”
- besede, ki vsebujejo najmanj tri a-je
- sedanjiške oblike glagola “delati”
- besede, ki vsebujejo najmanj 2 “lj”
- besede, ki vsebujejo dva zaporedna šumnika
- kratice iz najmanj treh velikih črk

Vendar..

- skoraj vsako orodje ima rahlo različno sintakso regularnih izrazov
- vsi ne podpirajo vseh predstavljenih operatorjev
- nekateri jih pa podpirajo še bistveno več

Pojavnice in različnice

- angleško *token* in *type*
- pojavnica: kar se pojavi v besedilu
- različnica: različne pojavnice v besedilu
- čeprav so pojavnice lahko karkoli, se ponavadi enačijo z besedami (včasih pa tudi z ločili)

Primeri

- Koliko pojavnic/različnic je v stavkih
  - Pika je prišla domov.
  - Pri surovem krompirju se barva spremeni zaradi fermentov, pri kuhanem pa zaradi oksidacije.
  - Kljub temu Novo mesto potrebuje novo vpadnico.
  - Gori na gori gori.

Problemi z različnicami

- Kdaj sta dve pojavnici dejansko različni?
  - velike in male črke, npr. *Novo, novo, NOVO, NoVo*
  - naglasna znamenja: jêsen/jesén/jesen
  - razlika v besedni obliki, vendar ne v lemi, npr. *miza, mize, mizi,…*
  - razlika v pomenu, vendar ne v besedni obliki oz. lemi, npr.
    "Ure so bile pokvarjene" proti *"Ure so bile poldne"*
    "Hotela je domov" proti *"Hotela ni več v mestu"*
- potrebna normalizacija pojavnic

Konkordance

- analiza na osnovi pojavnic
- konkordančno jedro z okoljem: levim in desnim sobesedilom
- ena najstarejših metod za analizo besedi (npr. Konkordanca Trubarjevega katekizma, 1983)
- v nasprotju s tiskanimi konkordancami sodobni konkordančniki omogočajo samo izpis želene besede oz. izraza
- dobimo primere uporabe: koristno za določanje pomena
- pri prevelikem število pojavitev nekateri konkordančniki omogočajo naključno sito
- koristno je tudi sortiranje po jedru ali sobesedilu

Frekvenčni seznami

- seznami različnic skupaj s številom pojavitev
- izguba konteksta
- pove lahko npr. katere besede so najbolj pogoste v korpusu (jeziku)

Najbolj pogoste leme v iKorpusu

N°	Hits	Atts	
1	4780	lemma	biti
2	3127	lemma	in
3	2563	lemma	v
4	1660	lemma	z
5	1584	lemma	na
6	1530	lemma	za
7	1286	lemma	ki
8	1037	lemma	se
9	1011	lemma	ta
10	790	lemma	da
11	691	lemma	sistem
12	651	lemma	on
13	623	lemma	tudi
14	613	lemma	pri
15	612	lemma	pa
16	589	lemma	lahko
17	560	lemma	podjetje
18	523	lemma	podatek
19	510	lemma	proces

- napogostejše so funkcijske
- polnopomenske besede vseeno nakazujejo področje, ki ga pokriva korpus
- splošen vtis o korpusu in njegovem besednem zakladu
- koristno kot pripomoček za izbiro posameznih besed za nadaljnjo analizo
- koristno tudi sortiranje (od spredaj ali od zadaj)

Še nekaj primerov

N°	Hits	Atts	Hit
1	9888	lemma	operacijski sistem
2	3196	lemma	informacijski sistem
3	903	lemma	datotečen sistem
4	640	lemma	računalniški sistem
5	514	lemma	nov sistem
6	385	lemma	posloven sistem
7	318	lemma	celoten sistem
8	230	lemma	inemiški sistem
9	226	lemma	različen sistem
10	208	lemma	velik sistem
11	182	lemma	obstoječ sistem
12	150	lemma	zmogljiv sistem
13	146	lemma	varnosten sistem
14	139	lemma	transakcijski sistem
15	121	lemma	nadzoren sistem
16	117	lemma	eksperten sistem
17	105	lemma	pomanjkljiv sistem
18	105	lemma	podoben sistem
19	97	lemma	navigacijski sistem

N°	Hits	Atts	Hit
1	9483	lemma	podpirati
2	1301	lemma	podati
3	703	lemma	podpreti
4	457	lemma	podajati
5	444	lemma	podpisati
6	396	lemma	podaljšati
7	338	lemma	podvojiti
8	216	lemma	podeliti
9	125	lemma	podražiti
10	123	lemma	podeljevati
11	123	lemma	podedovati
12	119	lemma	podariti
13	106	lemma	podreti
14	82	lemma	podleči
15	81	lemma	podkrepiti
16	79	lemma	podtaknti
17	66	lemma	podpisovati
18	62	lemma	podirati
19	61	lemma	podcenjevati

- kakšna sta bila iskalna izraza?
- čemu bi bili taki seznami koristni?

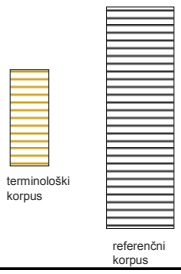
Statistične obdelave

- Z uporabo statističnih metod lahko odgovorimo na vprašanja, kot so:
  - katere besede najbolje opišejo neko besedilo?
  - katere besede najbolje razlikujejo dve besedili?
  - katere besede se najraje sopojavljajo z neko določeno besedo?
- večina teh metod primerja neko specifično besedišče s splošnim besediščem
- za vsako nalogo obstaja več konkurenčnih formul..

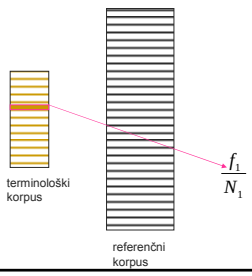
Ključne besede

- besede, ki najbolje opišejo neko besedilo (ali (pod)korpus)
- primerjamo število pojavitev vseh besed v našem besedilu s številom pojavitev teh besed v referenčnem korpusu
- število pojavitev delimo s številom besed v besedilu oz. ref. korpusu
- formula za "ključnost"
- opisno, če se npr. neka beseda pojavi v korpusu 1%, v besedilu pa 1.1%, ni ključna beseda, če pa 10%, pa je.

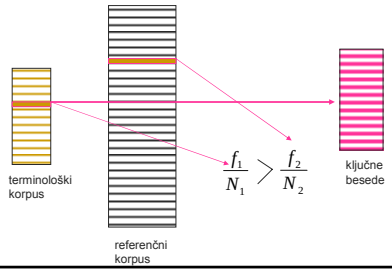
Luščenje ključnih besed



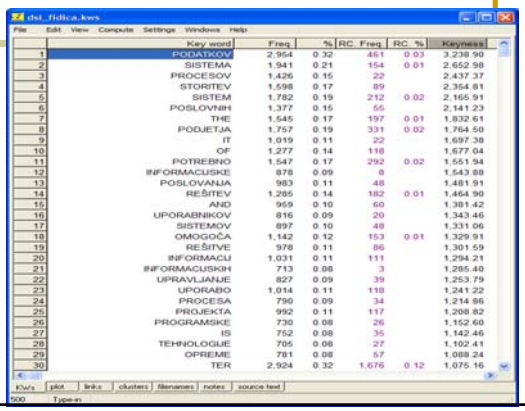
Luščenje ključnih besed



Luščenje ključnih besed



Primer iz Wordsmitha



	Key word	Freq.	%	RC Freq.	RC %	Keyness
1	PODATKOV	2,954	0.32	461	0.03	3,238.90
2	SISTEMA	1,941	0.21	154	0.01	2,652.98
3	PROCESOV	1,426	0.15	22		2,437.37
4	STORITEV	1,598	0.17	89		2,354.81
5	SISTEM	1,782	0.19	212	0.02	2,165.91
6	POSLOVNIH	1,377	0.15	55		2,141.23
7	THE	1,545	0.17	197	0.01	1,832.61
8	PODJETJA	1,757	0.19	331	0.02	1,764.50
9	IT	1,019	0.11	22		1,697.38
10	OF	1,277	0.14	118		1,677.04
11	POTREBNO	1,547	0.17	292	0.02	1,551.94
12	INFORMACIJSKE	878	0.09	8		1,543.88
13	POSLOVANJA	983	0.11	48		1,481.91
14	REŠITEV	1,285	0.14	182	0.01	1,464.90
15	AND	959	0.10	60		1,381.42
16	UPORABNIKOV	816	0.09	20		1,343.46
17	SISTEMOV	897	0.10	48		1,331.06
18	OMOGOČA	1,142	0.12	153	0.01	1,329.91
19	REŠITVE	978	0.11	86		1,301.59
20	INFORMACIJ	1,031	0.11	111		1,294.21
21	INFORMACIJSKIH	713	0.08	3		1,285.40
22	UPRAVLJANJE	827	0.09	39		1,253.79
23	UPORABO	1,014	0.11	118		1,241.22
24	PROCESA	790	0.09	34		1,214.86
25	PROJEKTA	992	0.11	117		1,208.82
26	PROGRAMSKE	730	0.08	26		1,152.60
27	IS	752	0.08	35		1,142.46
28	TEHNOLOGIJE	705	0.08	27		1,102.41
29	OPREME	781	0.08	57		1,088.24
30	TER	2,924	0.32	1,676	0.12	1,075.16

Luščenje “terminov”: TF-IDF

- iskanje podatkov (IR) – indeksiranje dokumentov
- namen: poiskati besede, ki naredijo dokument najbolj prepoznaven v množici in po katerih se najbolj razlikuje od vseh dokumentov v množici
- TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF

slovenski del JRC-Acquis / podkorpus besedil s področja jedrske energije

sevanju	0,19082	cepitve	0,05684
radiološkega	0,17864	nivoji	0,05684
dozimetrijo	0,17052	efektivno	0,05684
sivert	0,13804	medicine	0,05278
radionuklidov	0,13804	fuzije	0,05075
sevanja	0,13195	zaposlitvijo	0,04872
Dana	0,12992	termonuklearni	0,04872
Černobil	0,12180	študentov	0,04872
Izpostavljenost	0,12180	guvernerjev	0,04872
Jedrska	0,11368	prioritete	0,04872
dozo	0,09473	reaktorja	0,04872
prebivalstva	0,09256	jedrske	0,04872
sevanjem	0,08932	delodajalca	0,04669
ITER	0,08120	izpostavljenih	0,04601
Oddelkom	0,07308	ionizirajočemu	0,04466
inovativnosti	0,07308	ekvivalentno	0,04263
študente	0,07308	dosegljive	0,04060
izpostavljenosti	0,07308	ionizirajočega	0,04060
radioaktivne	0,06766	jedrskem	0,04060
SRS	0,06766	nuklearnih	0,04060
doza	0,06496	kontrolirana	0,04060
posameznike	0,06090	radiološki	0,04060
pooblaščenimi	0,05684		

Kolokacije

- statistično pogoste besedne zveze: nekatere besede družijo se rade
- idiomi, fraze, termini…
- več formul, ki primerjajo “naključno” porazdelitev sopojavitve besed z dejansko sopojavitvijo: MI, MI3, LL

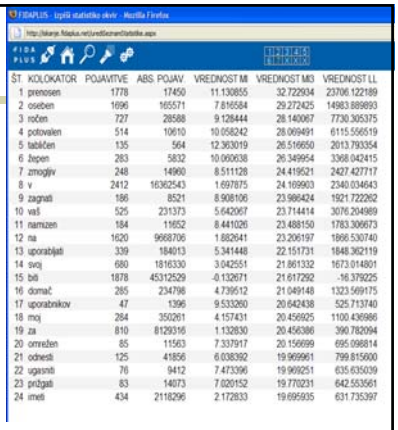
Naivni pristop

Query: IKORPUS; [word=".*"] [lemma="računalnik"]

N°	Hits	Atts	Hit
1	3400	lemma	oseben računalnik
2	3053	lemma	ročen računalnik
3	2983	lemma	v računalnik
4	2155	lemma	prenosen računalnik
5	1629	lemma	z računalnik
6	1439	lemma	. računalnik
7	1272	lemma	na računalnik
8	1219	lemma	biti računalnik
9	1031	lemma	žepen računalnik
10	890	lemma	namizen računalnik
11	708	lemma	za računalnik
12	494	lemma	svoj računalnik
13	418	lemma	drug računalnik
14	414	lemma	domač računalnik
15	408	lemma	omrežen računalnik
16	397	lemma	nov računalnik
17	388	lemma	ves računalnik
18	356	lemma	iz računalnik

"Kolokator + računalnik"
v FidaPLUS

Vrednosti MI, MI3 in LL



ŠT.	KOLOKATOR	POJAVITVE	ABS. POJAV.	VREDNOST MI	VREDNOST MI3	VREDNOST LL
1	prenosen	1778	17450	11.130855	32.722934	23706.122189
2	oseben	1696	165571	7.816584	29.272425	14983.889893
3	ročen	727	28588	9.126444	28.140067	7730.305375
4	potovalen	514	10610	10.058242	28.069491	6115.556519
5	tabličen	135	564	12.363019	26.516650	2013.793354
6	žepen	283	5832	10.060638	26.349954	3368.042415
7	zmogljiv	248	14960	8.511128	24.419521	2427.427717
8	v	2412	16362543	1.697875	24.169903	2340.034643
9	zagnati	186	8521	8.908106	23.986424	1921.722262
10	vaš	525	231373	5.642067	23.714414	3076.204989
11	namizen	184	11652	8.441026	23.488150	1783.306673
12	na	1620	9668706	1.882641	23.206197	1866.530740
13	uporabljati	339	184013	5.341448	22.151731	1848.362119
14	svoj	680	1816330	3.042551	21.861332	1673.014801
15	biti	1878	45312529	-0.132671	21.617292	-16.379225
16	domač	285	234798	4.739512	21.049148	1323.569175
17	uporabnikov	47	1396	9.533260	20.642438	525.713740
18	moj	264	350261	4.157431	20.456925	1100.436986
19	za	810	8129316	1.132830	20.456386	390.782094
20	omrežen	85	11563	7.337917	20.156699	695.096814
21	odnesti	125	41856	6.038392	19.969961	799.815600
22	ugasniti	76	9412	7.473396	19.969251	635.635039
23	prižgati	83	14073	7.020152	19.770231	642.553561
24	imeti	434	2118296	2.172833	19.695935	631.735397

Pa zaključimo

- kaj delajo konkordančniki
- regularni izrazi
- pojavnice in različnice
- kaj so konkordance in frekvenčni seznami
- in še nekaj statističnih metod:
  - ključnost in TF-IDF
  - kolokacije