



Content Analysis of *Institute for Information Studies'* Publications: Pilot Study

**Department of Information
Sciences
Faculty of Humanities and Social
Sciences**

**Petra Bago
Dina Crnec
Ph.D. Sanja Seljan
Ph.D. Hrvoje
Stančić**



CONTENTS

- INTRODUCTION
- *NooJ*
- PILOT STUDY –
Digitization
- RESEARCH RESULTS
- CONCLUSION
- REFERENCES



INTRODUCTION

- aim of the project:
 - language and content analysis of *Institute for Information Studies'* (Zavod za informacijske studije) publications
- gain knowledge into:
 - usage of terminology (frequency, collocations, n-grams etc.)
 - chronological appearance of terminology
 - penetration of cited resources (issuing date compared with date of citation)
- 19 publications 1990-2009 (cca 1 per year)

Publications (all in Croatian)

1. **Informacijske znanosti i znanje (1990)**
2. M. Tuđman, **Obavijest i znanje**
3. A. Stipčević, **Cenzura u knjižnicama**
4. S. Jelušić, **Struktura i organizacija knjižničnih sustava**
5. **Obrada jezika i prikaz znanja**
6. I. Maroević, **Uvod u muzeologiju**
7. T. Aparac-Gazivoda, **Teorijske osnove knjižnične znanosti**
8. J. Lasić-Lazić, **Znanje o znanju**
9. B. Tepeš, **Računarska lingvistika**
10. Z. Dovedan, **Formalni jezici: sintaksna analiza**
11. **Zbornik radova "Težakovi dani"**
12. **Modeli znanja i obrada prirodnog jezika**
13. D. Kovačević, J. Lasić-Lazić, J. Lovrinčević, **Školska knjižnica – korak dalje**
14. **Odabrana poglavlja iz organizacije znanja**
15. **Informacijske znanosti**
16. J. Lovrinčević, D. Kovačević, J. Lasić-Lazić, M. Banek Zorica, **Znanjem do znanja**
17. J. Lasić-Lazić, M. László, D. Boras, **Informacijsko čitanje**
18. S. Špiranec, M. Banek Zorica, **Informacijska pismenost: Teorijski okvir i polazišta**
19. H. Stančić, **Digitalizacija (2009)**

INTRODUCTION

- pilot study of the last issued publication
 - ready available in electronic form
 - published in 2009
- project aims to continue backwards
- publications in e-form + digitization & OCR





NooJ ⇒ NO Object

- a freeware linguistic engineering environment that includes large-coverage dictionaries and grammars
- a linguistic annotation system for corpus processing, parsing corpora in real time
- an information extraction system and a terminological extractor
- a *Machine Translation* development tool, a tool to teach linguistics and computational linguistics



NooJ ⇒ **NO Object**

- *NooJ*'s linguistic engine is multilingual
- modules for different languages available:
 - Arabic, Chinese, English, French, Hebrew, Spanish...
 - more modules under construction (i.e. Croatian)
- *NooJ* processes texts and corpora in 100+ file formats (i.e. HTML, PDF, MS Office, ASCII, even XML documents)



NooJ ⇒ NO Object

- it can process texts and corpora made of hundreds of text files
- *NooJ* dictionaries and grammars are extremely simple objects to build (with various tools for management)
- *NooJ* dictionaries represent all inflected and derived forms as one lexical entry ⇒ no more multiple independent lexical entries for derived forms!



NooJ ⇒ NO Object

➤ NooJ launches sophisticated queries over large corpora in order to produce various results (i.e. concordances, statistical analysis or information extraction)

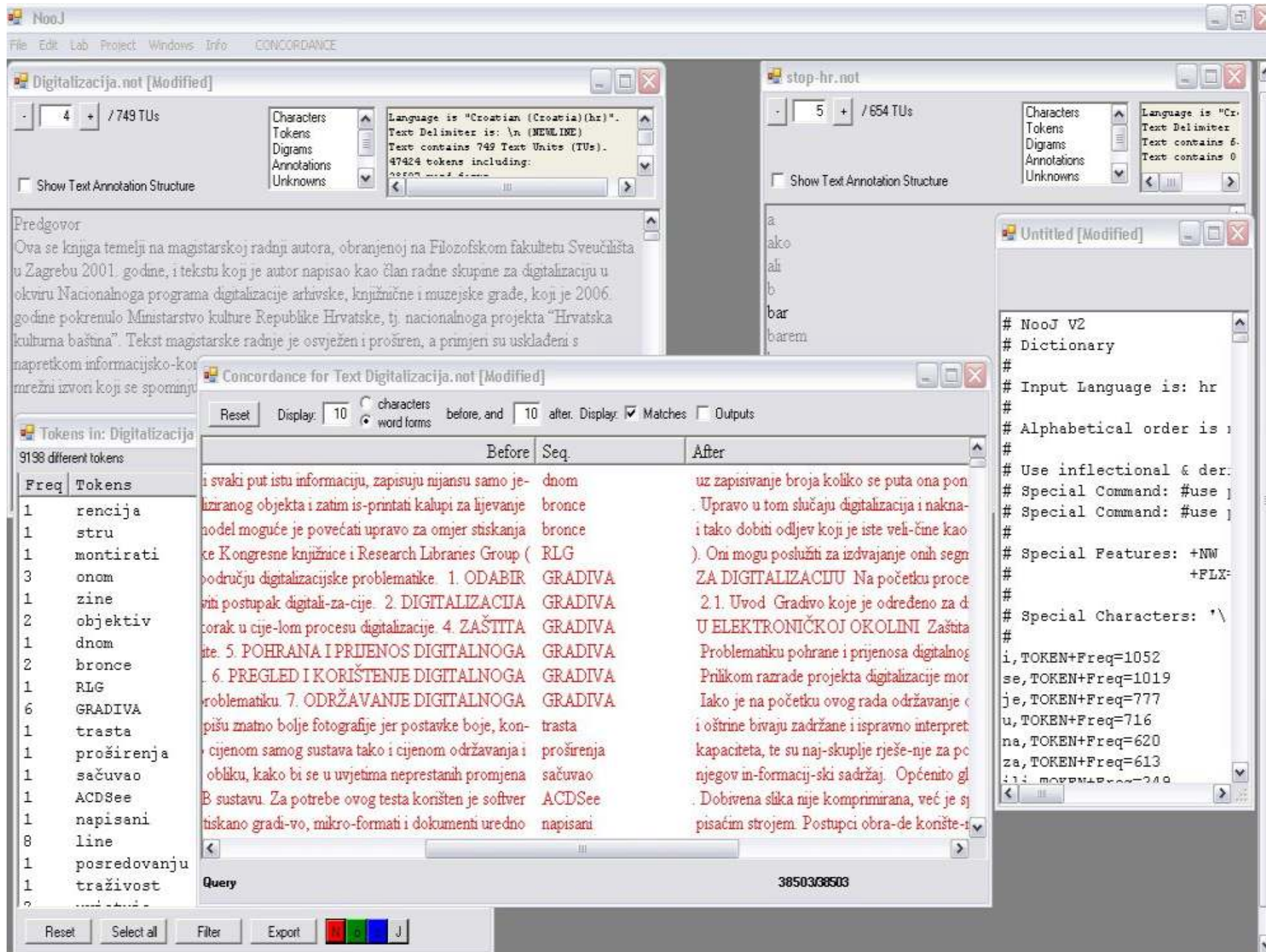
➤ **Zipf's law** ⇒ within a natural language corpus, the frequency of any word is inversely proportional to its rank in the frequency table

- the most frequent word will occur approximately twice as often as the second most frequent word
- 20% of words appear in 80% of text



NooJ ⇒ NO Object

- used as a main tool for creating lexical resources in this pilot study
- special attention given to:
 - tokens' frequency in the text
 - the analysis of n-grams (through bigrams)
 - English explanations occurring in the text
- the software, linguistic resources, manual, tutorials and reference papers at:
 - ⇒ <http://www.nooj4nlp.net>





PILOT STUDY – *Digitization*

- main problem:
 - no resources for Croatian ⇒ **manual extraction of terms (functional units)**
- list of stop-words for Croatian language (used for tokens/type)
- bigrams ⇒ retrieval of terms by a 10-word-search before and after functional units
- general curiosity (English) ⇒ *ETAOIN SHRDLU* represents an approximate order of frequency of the 12 most commonly used letters



PILOT STUDY – *Digitization*

➤ **AIM** ⇒ content analysis through a terminology database with a chronological overview and comparison of terms occurring in information sciences

➤ **results** of the research:

- half-automated extraction of stop-words and types without stop-words
- half-automated extraction of terms through bigrams by a 10-word-search before and after functional units

NooJ - [Tokens in: 19 Stancic, Digitalizacija - radno-mala slova]

File Edit Lab Project Windows Info

7414 different tokens

Freq	Tokens
1059	i
1018	se
778	u
771	je
626	za
624	na
349	ili
307	da
282	gradiiva
277	s
271	koji
248	su
219	može
212	od
202	koje
200	o
197	te
182	kao
167	bi
165	što

NooJ - [Untitled [Modified]]

File Edit Lab Project Windows Info DICTIONARY

```
# NooJ V2
# Dictionary
#
# Input Language is: hr
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' and properties' def:
# Special Command: #use properties.def
# Special Command: #use paradigms.nof
#
# Special Features: +NW (non-word) +FXC (frozen expression compo
#                   +FLX= (inflectional paradigm) +DRV= (derivat:
#
# Special Characters: '\ ' ' " ' + ' , ' # ' ' '
#
i, TOKEN+Freq=1059
se, TOKEN+Freq=1018
u, TOKEN+Freq=778
je, TOKEN+Freq=771
za, TOKEN+Freq=626
na, TOKEN+Freq=624
```

01- razlicnice i frekvencije.txt - Notepad

File Edit Format View Help

ñ	1059
se	1018
u	778
je	771
za	626
na	624
ili	349
da	307
gradiva	282
s	277
koji	271
su	248
može	219
od	212
koje	202
o	200
te	197
kao	182
bi	167
što	165
a	152
biti	151
gradivo	148
ne	144
će	142
zapisa	141
to	137
enq1	133

Microsoft Access

File Edit View Insert Tools Window Help



db1 : Database (Access 2000 file format)

Open Design New

- Objects
- Tables**
- Queries
- Forms
- Reports
- Pages
- Macros
- Modules
- Groups
- Favorites

- Create table in Design view
- Create table by using wizard
- Create table by entering data
- engl
- razlicnice**
- stop

14	od	212
15	koje	202
16	o	200
17	te	197
18	kao	182
19	bi	167
20	što	165
21	a	152
22	biti	151
23	gradivo	148
24	ne	144
25	će	142
26	zapisa	141
27	to	137
28	engl	133
29	mogu	119
30	podataka	116
31	digitalizacije	115

Record: 1 of 7414

db1 : Database (Access 2000)

razlicnice : Table

IDrazlicnica	razlicnica
1	i
2	se
3	u
4	je
5	za
6	na
7	ili
8	da
9	gradiva
10	s
11	koji
12	su
13	može
14	od
15	koje
16	o
17	te

Datasheet View

start



01-hr stop rijeci za bazu.txt - Notepad

File Edit Format View Help

```
"ID"      "stop"  
1         "ja"  
2         "mene"  
3         "me"  
4         "meni"  
5         "mi"  
6         "mnom"  
7         "mnome"  
8         "ti"  
9         "tebe"  
10        "te"  
11        "tebi"  
13        "tobom"  
14        "on"  
15        "njega"  
16        "ga"  
17        "njemu"  
18        "mu"  
19        "njem"  
20        "njim"  
21        "njime"  
22        "ona"  
23        "nje"  
24        "je"  
25        "njoj"  
26        "joj"  
27        "nju"  
28        "ju"  
29        "njom"
```

Reset Display: characters before, and after. Display: Matches Outputs
 word forms

Text	Before	Seq.	After
u svrhu namjernog zagušenja servera ((engl. undersampling) rezultira stepeničastim signalom (da imaju automatski uvlakač stranica (i cjelokupna problematika izrade sigurnosnih (se prenose po jedinici vremena (izvorno nastalog u elektroničkom obliku (gušći i nejasniji. podešavanje sjajnosti (javnog ključa upotrebom grube sile (prevelike količine podataka u međuspremniku (provjeriti identitet pošiljatelja. certifikacijska služba (kamere za snimanje pokretnih slika (postoji komprimiranje nepromjenjivim brojem bitova (gubicima: komprimiranje nepromjenjivim brojem bitova (fotoaparati sliku bilježi koristeći ccd (boje na zaslonima računala. cmyk (opcije za kontrolu ravnoteže boje (moraju se uvući u cam (autentičnosti i integriteta, * opisivanje sadržaja (sondiraju predmet fizičkim kontaktom. cmm (slično. 4.6. šifrirane omotnice šifrirane omotnice (mjeri u točkama po inču (inch – piksel po inču), dpi (koliko će točkica po inču (engl.		(distributed) denial of service – (d aliasing) te je ireverzibilan proces auto document feeder – adf). obično backup) kopija. šesto poglavlje, koje bitrate) – što je veći broj born digitally), kao početnog koraka brightness) i rezolucije može popraviti brute force), tj. načinom koji buffer) na prihvatnoj strani kanala ca – certifying authorities), služba koja camcorder). ovi se termini često cbr – constant bitrate encoding) i cbr – constant bitrate encoding) i charged coupled device) čip, koji cmyk – cyan, magenta, yellow, black colour balance), krivulje tonova (engl computer assisted manufacturing) program content labeling), * kontrola korištenja te coordinate measuring machine) može biti cryptographic envelopes, cryptolopes) nast dots per inch – dpi). odabir dots per inch – točke po dots-per-inch – dpi) skener

la u svrhu namjernog zagušenja servera (engl. (distributed) Den
nje (engl. undersampling) rezultira stepeničastim signalom (engl. a
ući da imaju automatski uvlačač stranica (engl. Auto Document Fee
osnim preslikama ili, pak, sigurnosnim duplikatima (engl. backup co
upna problematika izrade sigurnosnih (engl. backup) kopija. Šesto pog
žavanje sigurnosnih kopija (engl. backup). Glavna karakteristika ne
jedinici vremena (engl. bitrate) - što je veći broj bitova u seku
orno nastalog u elektroničkom obliku (engl. born digitally), kao poče
podešavanje sjajnosti (engl. brightness) i rezolucije može popraviti r
sljuča upotrebom grube sile (engl. brute force), tj. načinom koji is
velike količine podataka u međuspremniku (engl. buffer) na prihva
iti identitet pošiljatelja. Certifikacijska služba (engl. CA - cert
a snimanje pokretnih slika (engl. camcorder). Ovi se termini često
je nepromjenjivim brojem bitova (engl. CBR - Constant Bitrate En
primiranje nepromjenjivim brojem bitova (engl. CBR - Constant Bi
i fotoaparati sliku bilježi koristeći CCD (engl. Charged Coupled D
a zaslonima računala. CMYK (engl. CMYK - Cyan, Magenta, Yellow, bla
ije za kontrolu ravnoteže boje (engl. colour balance), krivulje
se uvući u CAM (engl. Computer Assisted Manufacturing) program
ovjera autentičnosti i integriteta, * opisivanje sadržaja (engl. c
iraju predmet fizičkim kontaktom. CMM (engl. Coordinate Measuring Mach
4.6. Šifrirane omotnice šifrirane omotnice (engl. cryptographic env
van pristup podacima (engl. Direct Access Storage - DAS). Za ovakve s
jeri u točkama po inču (engl. dots per inch - dpi). Odabir rezo
sel po inču), DPI (engl. dots per inch - točke po inču) i LPI (eng
liko će točaka po inču (engl. dots-per-inch - dpi) skener za bil
lučivosti (najmanje 600 dpi (engl. dpi - dots per inch, tj. broj toč
trebamo nabaviti pokretački program (engl. driver). Na Internetu čem
2.1.1.4. Rotacioni skeneri Rotacione skenerne (engl. drum scanner) već
ostrano (engl. simplex) ili obostrano (engl. duplex) skeniranje. Kod o
ve za upravljanje elektroničkim dokumentima (engl. EDMS - Electronic
za upravljanje elektroničkim dokumentima (engl. Electronic Docume
lektroničkim zapisima (spisovodstveni sustavi ili engl. Electroni
raju u zamišljeni kansasulu (engl. encapsulation) te spremaju. Ovaka

10 prije	izraz	10 posl
mnogih, neovlašteno preuzetih, računala u svrhu namjernog zagušenja servera (engl.	(Distributed) Denial of Service – (D)DOS), neželjene
orni signal. Nedovoljno brzo uzorkovanje (engl. undersampling) rezultira stepeničastim signalom (engl.	aliasing) te je ireverzibilan proces zbog prevelikog g
edajima ili fotokopirnim aparatima budući da imaju automatski uvlakač stranica (engl.	Auto Document Feeder – ADF). Obično se koriste :
zervnim ili pričuvnim kopijama, sigurnosnim preslikama ili, pak, sigurnosnim duplikatima (engl.	backup copy). "Sigurnosna kopija ... jest kopija poc
pohranu. Također se obrađuje i cjelokupna problematika izrade sigurnosnih (engl.	backup) kopija. Šesto poglavlje, koje se odnosi na j
sto koriste kao sustavi za izradu i održavanje sigurnosnih kopija (engl.	backup). Glavna karakteristika neizravnih sustava je
isi o broju bitova koji se prenose po jedinici vremena (engl.	bitrate) – što je veći broj bitova u sekundi, to je
igitaliziranoga gradiva, ali i onog izvorno nastalog u elektroničkom obliku (engl.	born digitally), kao početnog koraka u tom procesu.
veći što je tekst gušći i nejasniji. Podešavanje sjajnosti (engl.	brightness) i rezolucije može popraviti rezultate, ali
nog samo na osnovi poznavanja javnog ključa upotrebom grube sile (engl.	brute force), tj. načinom koji isprobava sve moguće
z uzrokovanja nedovoljne ili, pak, prevelike količine podataka u međuspremniku (engl.	buffer) na prihvatnoj strani kanala (dijagram 1). Nepr
ko bi primatelj podataka mogao provjeriti identitet pošiljatelja. Certifikacijska služba (engl.	CA – certifying authorities), služba koja dodjeljuje d
znih klasa te digitalne video kamere za snimanje pokretnih slika (engl.	camcorder). Ovi se termini često miješaju pa je nec
ži. U skladu s tim postoji komprimiranje nepromjenjivim brojem bitova (engl.	CBR – Constant Bitrate Encoding) i promjenjivim br
nskog i video gradiva s gubicima: komprimiranje nepromjenjivim brojem bitova (engl.	CBR – Constant Bitrate Encoding) i promjenjivim br
njesto na fotoosjetljivi film, digitalni fotoaparati sliku bilježi koristeći CCD (engl.	Charged Coupled Device) čip, koji je sastavljen od v
stav primarno koristi za prikaz boje na zaslonima računala. CMYK (engl.	CMYK – Cyan, Magenta, Yellow, black) sustav stv:
ogrami za skeniranje obično sadrže opcije za kontrolu ravnoteže boje (engl.	colour balance), krivulje tonova (engl. tonal curves),
jekti puste u proces printanja, moraju se uvući u CAM (engl.	Computer Assisted Manufacturing) program koji izr:
tekla eventualna nelegalna kopija, * provjera autentičnosti i integriteta, * opisivanje sadržaja (engl.	content labeling), * kontrola korištenja te * zaštita s
ekontakti. Kontaktni 3D skeneri sondiraju predmet fizičkim kontaktom. CMM (engl.	Coordinate Measuring Machine) može biti vrlo preci
avima, podatke o primatelju i slično. 4.6. Šifrirane omotnice Šifrirane omotnice (engl.	cryptographic envelopes, cryptolopes) nastale su k:
o im i samo ime govori, omogućuju izravan pristup podacima (engl.	Direct Access Storage – DAS). Za ovakve se susta
enirane slike. Razlučivost se obično mjeri u točkama po inču (engl.	dots per inch – dpi). Odabir rezolucije ovisi o potreb
šaka – PPI (engl. pixel per inch – piksel po inču), DPI (engl.	dots per inch – točke po inču) i LPI (engl. lines
zoluciju. Zapravo, mora se odrediti koliko će točaka po inču (engl.	dots-per-inch – dpi) skener zabilježiti i koliko će inf
ona mora biti skenirana u visokoj razlučivosti (najmanje 600 dpi (engl.	dpi – dots per inch, tj. broj točaka po kvadratnom in
ko stari čitač? Vjerojatno ne. Dakle, trebamo nabaviti pokretački program (engl.	driver). Na Internetu ćemo ga možda pronaći ako pr
ravo mediju za koji su namijenjeni. 2.2.1.1.4. Rotacioni skeneri Rotacione skenere (engl.	drum scanner) većinom rabe profesionalni studiji za
eneri imaju dvije glavne opcije – jednostrano (engl. simplex) ili obostrano (engl.	duplex) skeniranje. Kod obostranog skeniranja, prili
risteći danas već dohron razvijene sustave za upravljanje elektroničkim dokumentima (engl.	FDMS – Electronic Document Management System

Microsoft Access

File Edit View Insert Query Tools Window Help



db1 : Database (Access 2000 file format)

stoptxt : Select Query

```
SELECT razlicnice.IDrazlicnica, razlicnice.razlicnica  
FROM razlicnice, stop  
WHERE razlicnice.razlicnica=stop.stop;
```

Microsoft Access

File Edit View Insert Format Records

db1 : Database (Access 2000 file)

stoptxt : Select Query

IDrazlicnica	razlicnica
1	i
2	se
3	u
4	je
5	za
6	na
7	ili
8	da
10	s
11	koji
12	su
13	može
14	od
15	koje
16	o
17	te
18	kao

3	u
4	je
5	za
6	na
7	ili
8	da
10	s
11	koji
12	su
13	može
14	od
15	koje
16	o
17	te
18	kao
19	bi
20	što
21	a
22	biti
24	ne
25	će
27	to
29	mogu
32	kako

Record: 1 of 423

Microsoft Access

File Edit View Insert Query Tools Window Help



db1 : Database (Access 2000 file format)

Open Design New | X | [Lock] [Unlock] [Table] [Table] [Table] [Table] [Table]

razlicnice-stoptxt : Select Query

```
SELECT razlicnice.IDrazlicnica, razlicnice.razlicnica  
FROM razlicnice LEFT JOIN stop ON razlicnice.razlicnica=stop.stop  
WHERE stop.stop Is Null;
```

db1 : Database (Access 2000 fi

Open Design New

razlicnice-stoptxt : Select Q

IDrazlicnica	razlicnica
9	gradiva
23	gradivo
26	zapisa
28	engl
30	podataka
31	digitalizacije
34	medija
38	sustava
40	potrebno
44	kopija
46	sustav
49	digitalizaciju
50	prilikom
55	pohranu
57	video
59	slika
62	skeneri

20	engl
30	podataka
31	digitalizacije
34	medija
38	sustava
40	potrebno
44	kopija
46	sustav
49	digitalizaciju
50	prilikom
55	pohranu
57	video
59	slika
62	skeneri
64	slike
66	obliku
68	skeniranja
69	digitalni
71	kvalitete
72	slučaju
73	dokumenata
74	sustavi
75	zapis

Record: 1 of 6991

NooJ
File Edit Lab Project Windows Info

19 Stancic, Digitalizacija - radno-mala slova.not

487 / 752 TUs

Language is "Croatian (Croatia)(hr)".
Text Delimiter is: \n (NEWLINE)
Text contains 752 Text Units (TUs).
43989 tokens including:

Show Text Annotation Structure
 Characters
 Tokens
 Digrams
 Annotations
 Unknowns

sustav digitalnog potpisivanja sastoji se od tri dijela: generatora ključeva, funkcije potpisivanja i funkcije provjere. proces se odvija na sljedeći način. osoba koja želi potpisati neki digitalni dokument najprije pokreće generator ključeva, čime dobiva jedinstveni set javnog i privatnog ključa. zatim pokreće funkciju potpisivanja, koja kao ulazne vrijednosti ima digitalni dokument i tajni ključ, što rezultira digitalnim potpisom. tako potpisani dokument, uz dodatak javnog ključa, pošiljatelj putem komunikacijskog kanala dostavi primatelju ili ih javno objavi. primatelj dokumenta koji želi provjeriti autentičnost dokumenta mora pokrenuti funkciju provjere koja kao ulazne vrijednosti ima dokument, digitalni potpis i javni ključ. funkcija provjere rezultira priznavanjem ili nepriznavanjem izvornosti digitalnog potpisa (dijagram 5).

drugi, znatno učinkovitiji oblik dodavanja i provjere digitalnog potpisa zasniva se na funkciji raspršenja (engl. hash function). funkcijom raspršenja stvara se jedinstveni niz znakova za svaki dokument. taj niz znakova znatno je manji od dokumenta na temelju kojeg je stvoren te se naziva elektroničkim otiskom prsta. tada se e-otisak prsta šifira upotrebom kombinacije javnog i tajnog

Reoc. Digrams in Text: 19 Stancic, Digitaliza...

3889 reoccurring digrams / 26900

Freq	Digrams
2	digitalnih potvrda
2	definicija zapisa
2	constant bitrate
2	s medija
2	usb sučelja
2	digitalnog fotoaparata
2	gledajući s
2	ovog problema
2	kao nositelja
2	s originalom
2	na digitalnom
2	video kamere
2	od ponedjeljka
2	je stoga
2	se prema
2	dokumentima engl

Concordance for Text 19 Stancic, Digitalizacija - radno-mala slova.not

Reset Display: 10 characters before, and 10 after. Display: Matches Outputs

Text	Before	Seq.	After
dalje spajaju ili na pojačalo ili izravno na računalo putem uvijek skupo rješenje, prenosivi, neosjetljivi na vanjske utjecaje, spajanje putem	usb sučelja	usb sučelja	sljedeća slika prikazuje prednju i stražnju stranu pretpojačala. za digitalizaciju * udaljeni servisi za izradu sigurnosnih kopija – mrežna izrada pričuvnih kopija

Query 2/2

1. za digitalizaciju,DIGRAM+Freq=54 + proces digitalizacije,DIGRAM+Freq=14
2. za pohranu,DIGRAM+Freq=51
3. sigurnosnih kopija,DIGRAM+Freq=36 + sigurnosne kopije,DIGRAM+Freq=26 + sigurnosna kopija,DIGRAM+Freq=21
4. digitalnoga gradiva,DIGRAM+Freq=29
5. brojem bitova,DIGRAM+Freq=23
6. digitaliziranoga gradiva,DIGRAM+Freq=21
7. gradivo koje,DIGRAM+Freq=21 + gradiva u,DIGRAM+Freq=20 + se gradivo,DIGRAM+Freq=20 + gradiva koje,DIGRAM
8. u elektroničkom,DIGRAM+Freq=19 + elektroničkom obliku,DIGRAM+Freq=18
9. i video,DIGRAM+Freq=18
10. za skeniranje,DIGRAM+Freq=16
11. slikovnoga gradiva,DIGRAM+Freq=16
12. sustava za,DIGRAM+Freq=16
13. za arhiviranje,DIGRAM+Freq=14
14. elektroničkoga gradiva,DIGRAM+Freq=14
15. u boji,DIGRAM+Freq=13
16. zahtjev za,DIGRAM+Freq=13 + zahtjeva za,DIGRAM+Freq=12
17. za izradu,DIGRAM+Freq=13
18. s gubicima,DIGRAM+Freq=13 + bez gubitaka,DIGRAM+Freq=12
19. u digitalnom,DIGRAM+Freq=12
20. zvučnoga gradiva,DIGRAM+Freq=12
21. javni ključ,DIGRAM+Freq=12
vrsta gradiva,DIGRAM+Freq=11
sustavi za,DIGRAM+Freq=11
pružatelja usluga,DIGRAM+Freq=11
zvučnog signala,DIGRAM+Freq=11
prilikom digitalizacije,DIGRAM+Freq=10
na računaru,DIGRAM+Freq=10
uređaji za,DIGRAM+Freq=10
procesa digitalizacije,DIGRAM+Freq=10
magnetske trake,DIGRAM+Freq=10

šifriranja javnim,DIGRAM+Freq=2
radnog procesa,DIGRAM+Freq=2
medija no,DIGRAM+Freq=2
skeniranih dokumenata,DIGRAM+Freq=2
distribuiranja te,DIGRAM+Freq=2
stručnjaci za,DIGRAM+Freq=2
građe radi,DIGRAM+Freq=2
neki skeneri,DIGRAM+Freq=2
21. profesionalna oprema,DIGRAM+Freq=2
20. projekti digitalizacije,DIGRAM+Freq=2
19. svrhu očuvanja,DIGRAM+Freq=2
18. arhivima knjižnicama,DIGRAM+Freq=2
17. memorijsku karticu,DIGRAM+Freq=2
16. stranice knjige,DIGRAM+Freq=2
15. komprimirani zapis,DIGRAM+Freq=2
14. potpunost autentičnost,DIGRAM+Freq=2
13. arhiviranje dokumenata,DIGRAM+Freq=2
12. rezolucije i,DIGRAM+Freq=2
11. skenirati u,DIGRAM+Freq=2
10. problem konverzije,DIGRAM+Freq=2
9. d skenerima,DIGRAM+Freq=2
8. potrebe arhiviranja,DIGRAM+Freq=2
7. digitalnih potvrda,DIGRAM+Freq=2
6. s medija,DIGRAM+Freq=2
5. usb sučelja,DIGRAM+Freq=2
4. digitalnog fotoaparata,DIGRAM+Freq=2
3. kao nositelja,DIGRAM+Freq=2
2. na digitalnom,DIGRAM+Freq=2
1. video kamere,DIGRAM+Freq=2

Nacionalni program digitalizacije arhivske, knjižnične i muzejske građe			
Digitalizacija gradiva			
Postupak određivanja prioriteta za digitalizaciju			1. Digitalne video kamere za snimanje pokretnih slika
Odabir prave tehnologije za digitalizaciju			2. Automatski okidati snimku na digitalnom fotoaparatu
Stupanj u procesu digitalizacije			Usporedba žarišne duljine objektiva na digitalnom fotoaparatu s onom kod analognog
Proces odabira gradiva koje će se digitalizirati			3. Očuvanje fizičkog objekta kao nositelja informacije
Uređaji za digitalizaciju			Medij kao nositelj informacije
Digitalizacijska problematika			4. Konkretna kombinacija digitalnog fotoaparata i objektiva s klasičnog aparata
Odabir gradiva za digitalizaciju			Digitalizacija uporabom digitalnog fotoaparata rezultira slikom izvornika koja se poslije
Stručnjaci za digitalizaciju			5. Spajanje putem usb sučelja
Odabir dokumenata za digitalizaciju			6. Migracija nije samo prebacivanje s medija na neki novi medij
Smjernice za odabir građe za digitalizaciju			Premašivanje određenog broja grešaka prilikom čitanja zapisa s medija
Gradivo predloženo za digitalizaciju			7. Izdavanje digitalnih certifikata, tj. digitalnih potvrda kojima se dokazuje identitet
Oprema za digitalizaciju			Problem digitalnih potvrda o identitetu
Studio za digitalizaciju			8. Model kodiranja entropije za potrebe arhiviranja slikovnog gradiva
Mogućnosti prijevoza gradiva za digitalizaciju			9. Skeniranje 3d skenerima
Sustav za digitalizaciju			10. Problem konverzije zapisa
Posebni skeneri za digitalizaciju mikrofilma ili mikrofiša			11. 35 mm gradivo mora se skenirati u vrlo visokoj rezoluciji
Profesionalni studiji za digitalizaciju			Skenirati u boji
Skeneri praktični za digitalizaciju gradiva velikog forma			12. Potrebna rezolucija i dubina boje u odnosu na vrstu izvornika
Digitalizacija na terenu			Slika više rezolucije
Digitalizacija uvezanih dokumenata			13. Arhiviranje dokumenata
Uređaji za digitalizaciju zvuka			14. zadržati vjerodostojnost, potpunost, autentičnost i dovoljno konteksta migriranih za
Digitalizacija vinilskih ploča			15. Komprimirani zapis na zahtjevnijim dionicama
Kvaliteta digitaliziranih snimaka			Rezultirajući komprimirani zapis
Digitalizacija filma i videa			16. Istovremeno se digitaliziraju obje otvorene stranice knjige
Odabir tehnika i uređaj za digitalizaciju			17. Na memorijsku karticu mogu se snimati stotine snimaka
Različite tehnike za digitalizaciju filmskih materijala			Pohrana fotografija na memorijsku karticu u komprimiranom obliku
Korištenje projektor			18. Djelatnici u arhivima, knjižnicama, muzejima, drugim informacijsko-dokumentacijsk
Digitalizacija signala			19. Arhiviranje u svrhu očuvanja građe u digitalnom obliku
Uređaj za digitalizaciju filmskog gradiva			Djelovanje digitalizacijom u svrhu očuvanja nacionalnog identiteta i dostojnog predstavl
			20. Mnogi projekti digitalizacije bave se digitalizacijom tiskanih dokumenata
			21. Profesionalna oprema uz potpunu automatiku omogućuje i ručne postavke
			Profesionalna oprema se u pravilu izvodi u slr



CONCLUSION

- lack of resources for Croatian language
 - higher level of automation possible
 - present situation ⇒ a time consuming process
- pilot study resulted with defined analysis procedures for further research
- first results of a broader study
 - environment of tokens aiming for term extraction
- planning the research process for all 19 publications



➔ all publications of the
Institute for Information Studies
(Zavod za informacijske studije)
available online:

<http://infoz.ffzg.hr/bookstore/>



REFERENCES

- Delač, Davor et al. *TermeX: A Tool for Collocation Extraction*. Faculty of Electrical Engineering and Computing, University of Zagreb, 2009
- Silberztein, Max. NooJ.
<http://www.nooj4nlp.net>
- Stančić, Hrvoje. *Digitization*. Institute for Information Studies, Zagreb, 2009
- Zipf Wiki.
http://en.wikipedia.org/wiki/Zipf's_law

INFuture2009: Digital Information and Knowledge Sharing

➤ 4-6 November 2009

➤ Zagreb,
*Hotel Palace &
Faculty of Humanities
and Social Sciences*

➤ You are all
welcome!

<http://infoz.ffzg.hr/INFuture>

INFuture 2009
The Future of Information Sciences
4-6 November 2009
Zagreb, Croatia

Home | **Conference** | Venue | Papers | INFuture 2007

Conference
Conference
Registration

Venue
Travel
Accommodation
Visa regulations
About Croatia
About Zagreb

Papers
Important dates
Paper submission guidelines
Conference Proceedings

Conference

Department of Information Sciences
Faculty of Humanities and Social Sciences, Zagreb, Croatia

ORGANISE
2nd International Conference
The Future of Information Sciences (INFuture)

INFuture2009: "Digital Resources and Knowledge Sharing"

4-6 November 2009

<http://infoz.ffzg.hr/INFuture>, infuture@infoz.ffzg.hr

FIRST CALL FOR PAPERS

INFuture 2009: Digital Resources and Knowledge Sharing is the second in a series of INFUTURE conferences focusing on creation, sharing and reuse of digital resources. The objective of the conference is to provide a platform for discussing both theoretical and practical issues.

TOPICS (include but not limited to):

- Virtual Environment in Education
- Using Open-Source Solutions in Cultural Heritage
- Knowledge Management
- Using Information Resources in Research, Education and Presentation
- Digitization and Preservation
- Language Technologies
- e-Services, e-Government and Business Applications
- Special Session: Doctoral Colloquium

THANK YOU!

**Content Analysis of
Institute for Information Studies'
Publications:
Pilot Study**

**Department of Information
Sciences
Faculty of Humanities and Social
Sciences**

**Petra Bago
Dina Crnec
Ph.D. Sanja Seljan
Ph.D. Hrvoje
Stančić**