Positioning: ervaring met testcorpus 1

Ellen Rusman
Jan van Bruggen
April 2005

OpenUniversiteitNederland

Apencorpus

- Documentselectie
- Splitsen
- Opschonen
  - spelfouten
  - redundante teksten
  - Identieke teksten
  - HTML-code
  - diacritische tekens
  - plaatjes
  - witte ruimte
- Aantal documenten

Tools: Concordance

Concordance - Aap01Bonobo.txtPlus247MoreFiles.txt.Concordance

File Text Search Edit Headwords Contexts View Tools Help

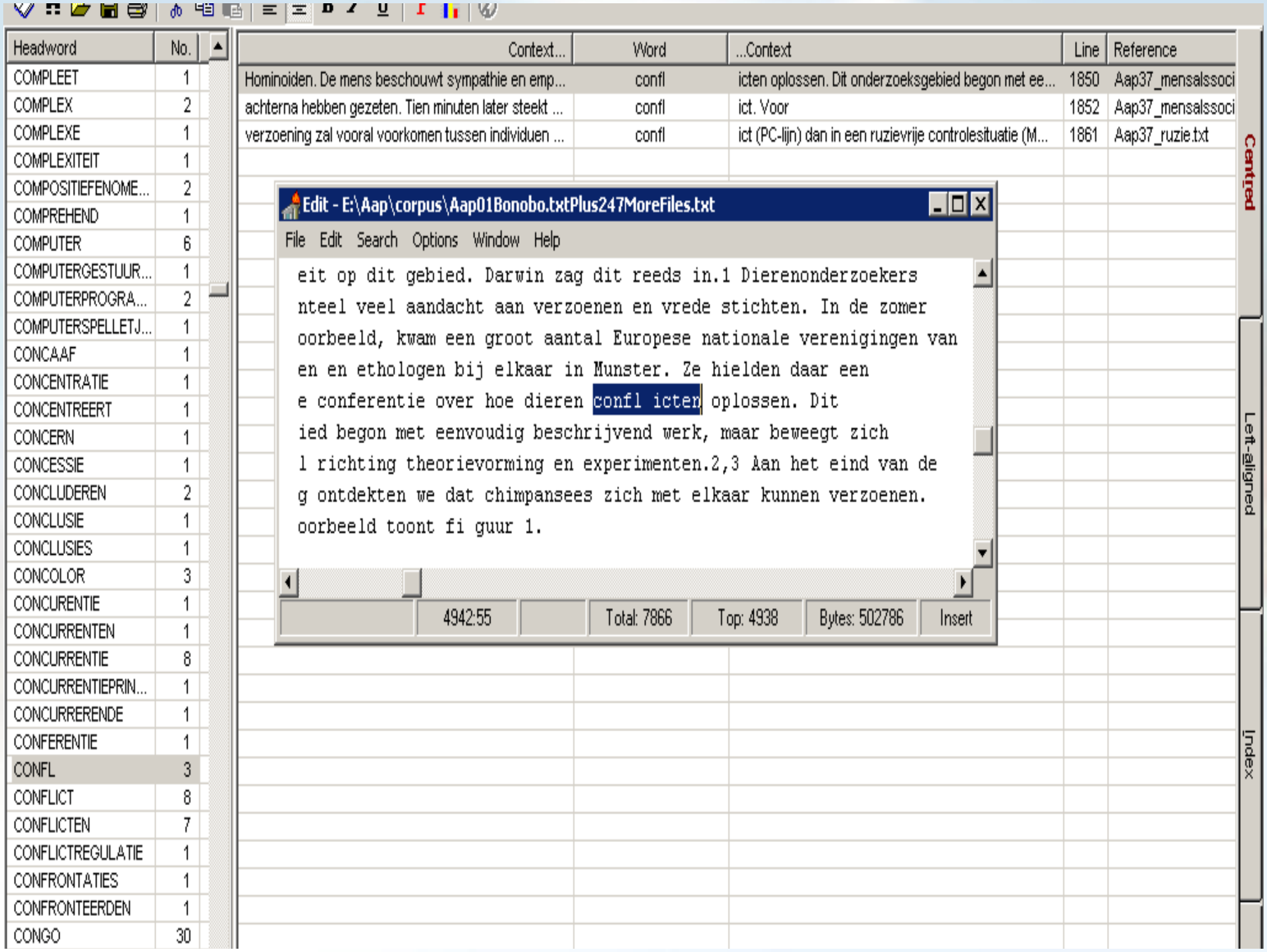
Headword	No.
SULU-EILANDEN	1
SUMATERA	8
SUMATRA	28
SUMATRAANSE	5
SUMULEREND	1
SUNDA-EILANDEN	1
SUPERFAMILIE	8
SUPERFAMILIES	3
SUPPLY	1
SURINAME	2
SUSHI	4
SUSHIMEESTER	2
SUSSEN	1
SUSSEND	1
SYLVANUS	4
SYMBOLICUS	1
SYMBOLISCHE	2
SYMBOLISEERT	1
SYMBOOL	4
SYMBOOLHANTERING	1
SYMPATHETIC	1
SYMPATHIE	3
SYMPHALANGUS	1

Context...	Word	...Context	Li...	Reference
Bekende soorten zijn: de zwarte siamang of imbau (S...	Sumatera	en Malakka, bij welke de tweede en de derde teen vo...	600	Aap12Gibbons.txt
Bekende soorten zijn: de zwarte siamang of imbau (S...	Sumatera	, de grijze wau-wau, oewa of zilvergibbon (Hylobates...	600	Aap12Gibbons.txt
Bekende soorten zijn: de zwarte siamang of imbau (S...	Sumatera	, met witte handen en voeten. Gibbons worden in hun ...	600	Aap12Gibbons.txt
De magot is de enige Afrikaanse soort. De resusaap i...	Sumatera	, de Java-aap of monjet (M. fascicularis) van Birma tot...	623	Aap12Makaken.txt
Orang-oetans bewonen de wouden van Noord-Sumat...	Sumatera	en Borneo zijn uiterlijk vrij goed van elkaar te ondersc...	635	Aap12OrangOetan.txt
Orang-oetans bewonen de wouden van Noord-Sumat...	Sumatera	) zeer teruggelopen; men tracht in beslag genomen die...	635	Aap12OrangOetan.txt
Orang-oetans bewonen de wouden van Noord-Sumat...	Sumatera	en Borneo, wat niet altijd met succes bekroond is. Slu...	635	Aap12OrangOetan.txt
In dierentuinen zijn orang-oetans goed te houden; de d...	Sumatera	en Borneo gescheiden te fokken, verzekert de raszui...	636	Aap12OrangOetan.txt

Centred

Left-aligned

OpenUniversiteitNederland


Headword
No.
COMPLEET 1
COMPLEX 2
COMPLEXE 1
COMPLEXITEIT 1
COMPOSITIEFENOME... 2
COMPREHEND 1
COMPUTER 6
COMPUTERGESTUUR... 1
COMPUTERPROGRA... 2
COMPUTERSPELLETJ... 1
CONCAAF 1
CONCENTRATIE 1
CONCENTREERT 1
CONCERN 1
CONCESSIE 1
CONCLUDEREN 2
CONCLUSIE 1
CONCLUSIES 1
CONCOLOR 3
CONCURENTIE 1
CONCURRENTEN 1
CONCURRENTIE 8
CONCURRENTIEPRIN... 1
CONCURRERENDE 1
CONFERENTIE 1
CONFL 3
CONFLICT 8
CONFLICTEN 7
CONFLICTREGULATIE 1
CONFRONTATIES 1
CONFRONTEERDEN 1
CONGO 30
Context...
Word
...Context
Line
Reference
Hominoiden. De mens beschouwt sympathie en emp... confl icten oplossen. Dit onderzoeksgebied begon met ee... 1850 Aap37_mensalssoci
achterna hebben gezeten. Tien minuten later steekt ... confl ict. Voor 1852 Aap37_mensalssoci
verzoening zal vooral voorkomen tussen individuen ... confl ict (PC-lijn) dan in een ruzievrije controlesituatie (M... 1861 Aap37_ruzie.txt
Edit - E:\Aap\corpus\Aap01Bonobo.txtPlus247MoreFiles.txt
File Edit Search Options Window Help
eit op dit gebied. Darwin zag dit reeds in.1 Dierenonderzoekers
nteel veel aandacht aan verzoenen en vrede stichten. In de zomer
oorbeeld, kwam een groot aantal Europese nationale verenigingen van
en en ethologen bij elkaar in Munster. Ze hielden daar een
e conferentie over hoe dieren confl icten oplossen. Dit
ied begon met eenvoudig beschrijvend werk, maar beweegt zich
l richting theorievorming en experimenten.2,3 Aan het eind van de
g ontdekten we dat chimpansees zich met elkaar kunnen verzoenen.
oorbeeld toont fi guur 1.
4942:55
Total: 7866
Top: 4938
Bytes: 502786
Insert
Centred
Left-aligned
Index

TextStat

TextSTAT - Aapcorpus.crp

Corpus Export Taal Codering ?

Corpus | Woordvormen | Concordantie | Citaat

Woordvorm	Frequentie
grond	104
5	104
één	103
zij	103
Bij	102
vacht	101
eerste	101
Afrika	99
lange	98
lang	97
komt	96
meestal	96
mannen	96
ander	96
Ook	96
echter	96
sociale	95
erg	94
net	93
gaat	93
nu	93
gorilla's	93
na	92
daar	91
tegen	90
tijd	90

Frequentie / opties

- (•) sorteren op frequentie
- ( ) sorteren op alfabet
- ( ) retrograde sorteren

--- min. frequentie

--- max. frequentie

[ ] hoofdletters negeren (A=a)

OF zoek frequentie voor woorden met de string:

Frequentielijst

12902 woordvormen/types (114111 woorden/tokens in corpus) | 8 bestanden | 725434 bytes

Stoppen en stemmen

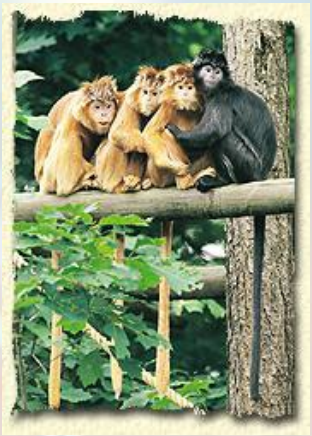
- Stoplijst
  - Methodiek
  - Apencorpus
  - Oracle+
  - Volkskrant
- Stemmen

Analyses

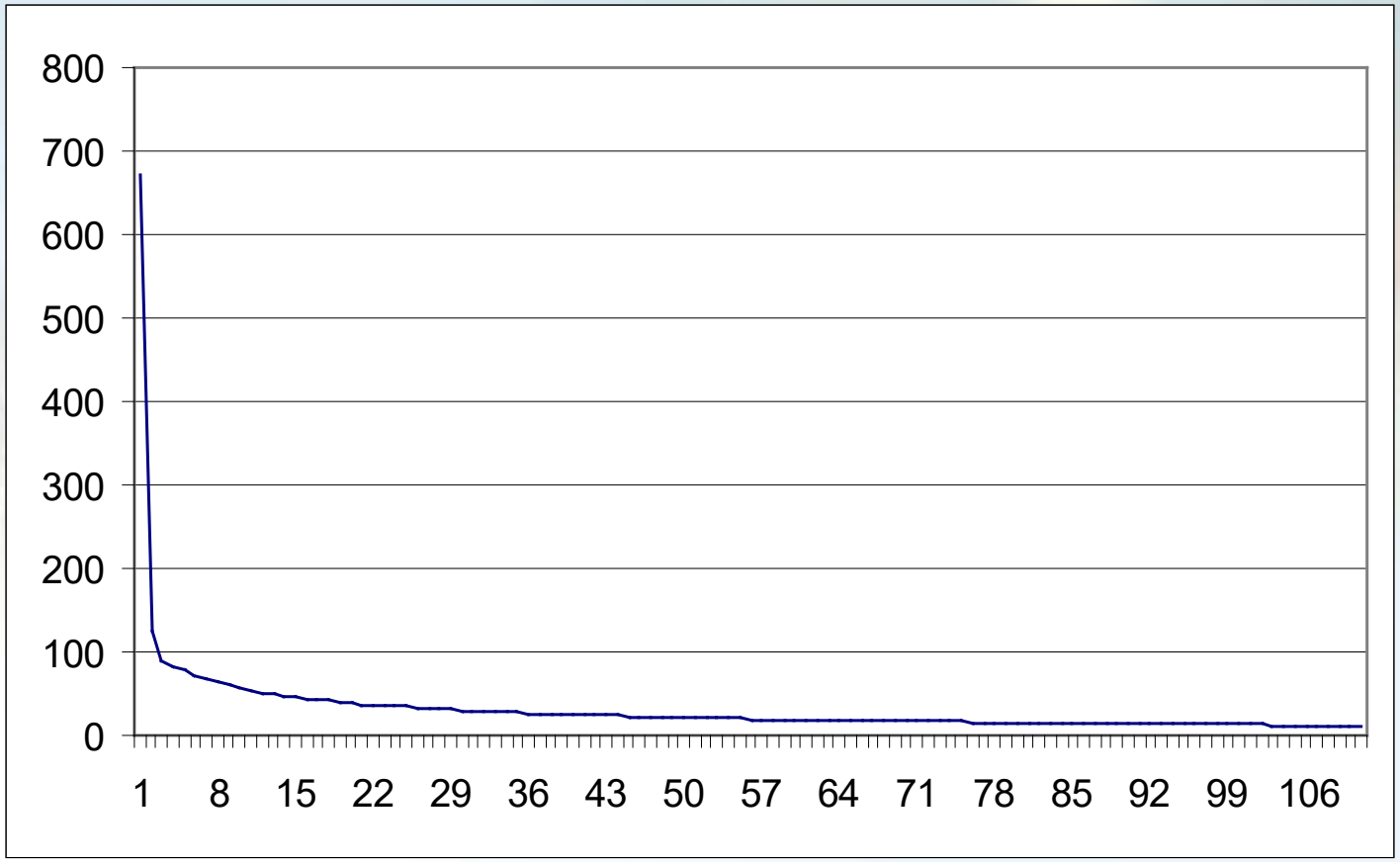
- Parameters uitproberen
  - SVD
  - Queries
- Interne consistentie
  - documenten over soort X
- Validiteit
  - documentcorrelaties
  - vergelijking met beoordelaars

Resultaten

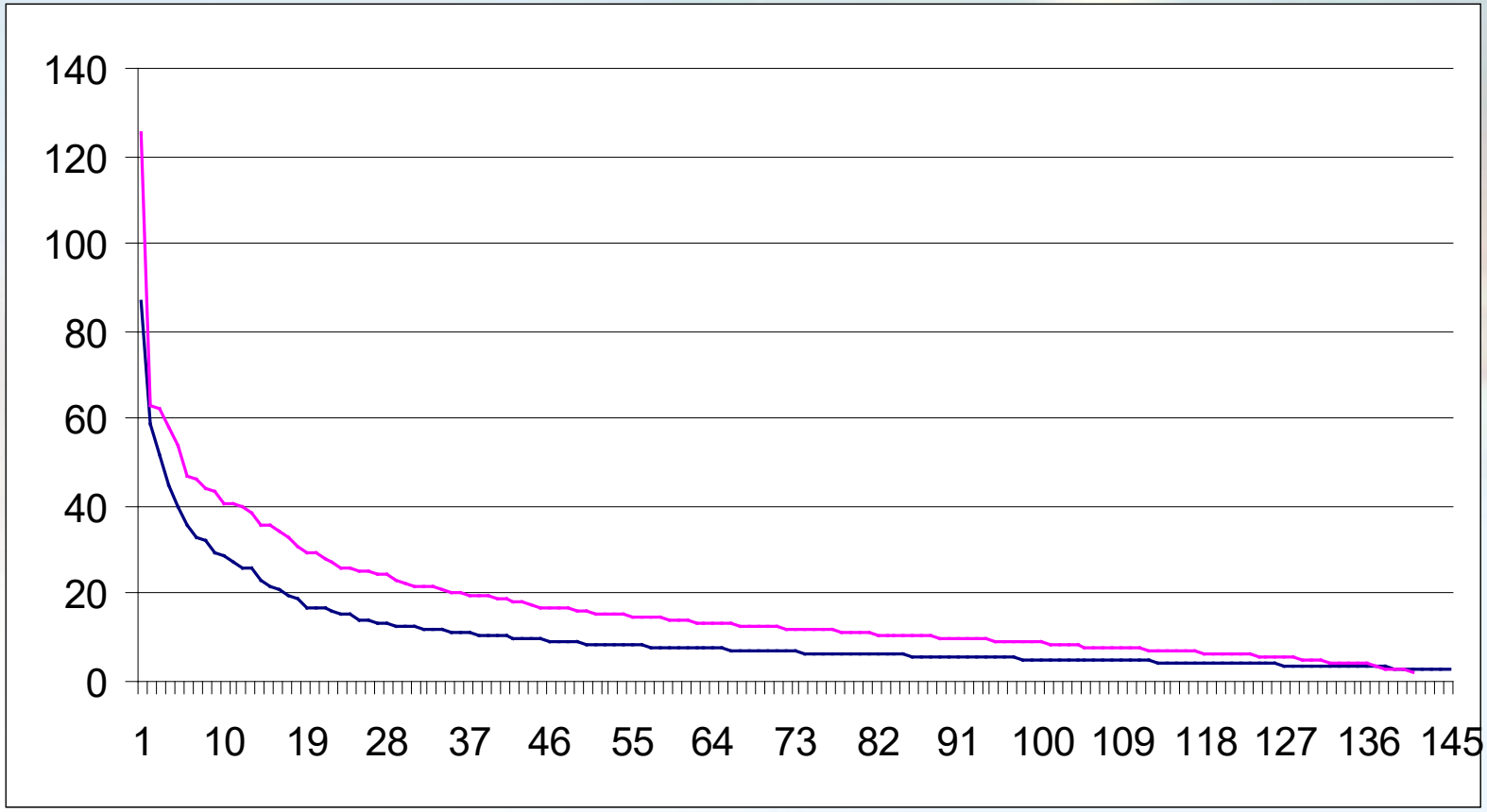
- Eigenschappen corpus
  - aantal documenten
  - aantal termen
  - aantal dimensies
    - hele corpus
    - gehalveerde corpus
- Interne consistentie
  - voorbeeldquery



Corpuseigenschappen



Corpuseigenschappen



	A	B	C	D	E	F	G	H	I	J	K	L
1		docnr										
2			4	20	44	60	72	144	162	216	217	
3	G	4	1.00	0.75	0.74	0.61	0.69	0.84	0.53	0.74	0.72	
4	G	20	0.75	1.00	0.72	0.56	0.63	0.81		0.72	0.56	
5	G	29	0.54	0.56	0.50		0.51	0.61		0.50	0.49	
6	G	44	0.74	0.72	1.00	0.62	0.64	0.76		1.00	0.64	
7	nieuwe aap	47	0.59	0.50	0.58		0.56	0.53		0.58	0.55	
8	nieuwe aap	48	0.43		0.43					0.43		
9	doodshoofd	53		0.42								
10	G	60	0.61	0.56	0.62	1.00	0.46	0.57	0.50	0.62	0.49	
11	G	72	0.69	0.63	0.64	0.46	1.00	0.72		0.64	0.57	
12	G	88	0.72	0.48	0.66	0.63	0.56	0.64		0.66	0.53	
13	mensapen	110	0.44		0.42		0.40			0.42		
14	G	119	0.82	0.56	0.69	0.53	0.65	0.72		0.69	0.62	
15	G	144	0.84	0.81	0.76	0.57	0.72	1.00		0.76	0.74	
16	G	148	0.83	0.67	0.74	0.52	0.70	0.80		0.75	0.78	
17	G	162				0.50			1.00			
18	G	165	0.60	0.52	0.69	0.54	0.53	0.59	0.59	0.69	0.43	
19	G	168	0.75	0.63	0.73	0.58	0.66	0.75		0.73	0.69	
20	G	170			0.40	0.42	0.41		0.76	0.40		
21	G	174							0.65			
22	doodshoofd	195		0.45								
23	G	214	0.82	0.64	0.70	0.52	0.63	0.75		0.70	0.63	
24	G	215	0.52	0.47	0.52	0.40	0.48	0.57		0.52	0.47	
25	G	216	0.74	0.72	1.00	0.62	0.64	0.76		1.00	0.64	
26	G	217	0.72	0.56	0.64	0.49	0.57	0.74		0.64	1.00	
27	mensapen	244	0.77	0.56	0.64	0.53	0.63	0.66		0.64	0.72	
28	orangoetan	257							0.43			
29	mensapen	284			0.41					0.41		
30												
31												

Conclusies

- Opbouw corpus
  - Minimale omvang
  - Opschonen en tools
- Stoppen
  - Werkbare methode gevonden (?)
- Stemmen
  - Hier niet gedaan, maar …
- Parameters
- Interne consistentie