

Intro til design og brug af korpora

Jørg Asmussen
ja@dsl.dk

Det Danske Sprog- og Litteraturselskab
www.dsl.dk

Intro til

design og brug
korpuslingvistik
af korpora

Jørg Asmussen
ja@dsl.dk

Det Danske Sprog- og Litteraturselskab
www.dsl.dk

Intro til

design og brug
korpuslingvistik
af korpora

Jørg Asmussen
ja@dsl.dk

Det Danske Sprog- og Litteraturselskab
et reklameindslag...

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

- Tekstudgivelser
- Sproghistoriske fremstillinger
- Bibliografier
- Ordbøger og sprogteknologi

Hvad er DSL?



Det Danske Sprog- og Litteraturselskab — Det Danske Sprog- og Litteraturselskab

http://dsl.dk/

Opslag ▾ DSL ▾ Lingvistik ▾ Konferencer ▾ Info ▾ Mac ▾

Sideoversigt ▾ Tilgængelighed ▾ Nyhedsarkiv ▾ Om DSL ▾ Kontakt

DSL
DET DANSKE SPROG- OG
LITTERATURSELSKAB

du er her: forside

Søg i dsl.dk: Søg

- ♥ Tekstudgivelser
- ♥ Ordbøger og sprogteknologi
- ♥ Sproghistoriske fremstillinger
- ♥ Bibliografier

DSL's digitale udgivelser

- Arkiv for Dansk Litteratur
- Diplomatarium Danicum
- Korpus 2000
- Ordbog over det danske Sprog
- Potentielle Klassikere
- Studér Middelalder på Nettet

Det Danske Sprog- og Litteraturselskab (DSL) udgiver og dokumenterer dansk sprog og litteratur fra de ældste tider til i dag – i bogform og på nettet. Selskabet beskæftiger ca. 30 videnskabelige medarbejdere og en væsentligt større kreds af medlemmer, som fører tilsyn med udgivelserne og udgør selskabets øverste myndighed. Selskabet blev stiftet i 1911 og modtager støtte fra Kulturministeriet, Carlsbergfondet og en række andre fonde. [Læs mere om DSL](#)

www.dsl.dk

Sproghistoriske fremstillinger

Det Danske Sprog- og Litteraturselskabs udgivelsesvirksomhed i form af tekstkritiske udgaver af historiske og litterære kilder og af ordbøger over forskellige sprogtrin og forfatterskaber leverer et fremragende og nødvendigt materiale for forskere til undersøgelse af det danske sprogs historie. Også inden for Selskabets egne rammer har disse kilder været udnyttet dels til samlede fremstillinger af det danske sprogs historie, dels til studier af sproghistoriske delemler, og i de seneste år er denne aktivitet intensiveret. [Læs mere om sproghistoriske fremstillinger](#)



Ordbøger og sprogteknologi

At udarbejde danske ordbøger på videnskabeligt grundlag har været en hovedopgave for Det Danske Sprog- og Litteraturselskab siden grundlæggelsen. Selskabet har udgivet flere store nationalordbøger og ordbøger over ældre sprogtrin og enkeltforfatterskaber. I de senere år har aktiviteten bredt sig til også at omfatte tosprogsordbøger mellem dansk og andre nordiske sprog samt netordbøger og tekstkorpusser. [Læs mere om ordbøger og sprogteknologi](#)

Seneste udgivelse



Herman Bang
Vekslende Themaer

I årene 1879–84 skrev den unge Herman Bang 210 søndagskronikker, såkaldte feuilletoner, under fællestitlen »Vekslende Themaer«.

Udgivet med noter og efterskrift af Sten Rasmussen.

4 bind, 1796 sider.
Kr. 495,00.

Udkom februar 2007.
ISBN 978-87-7876-465-2

Se flere af DSL's udgivelser

Program

1. Tekstkorpora og referencekorpora

2. Korpussammensætning

3. Korpusopmærkning

4. Korpusundersøgelser

5. Fremtiden

Korpuslingvistik

1. Sprogbeskrivelse på baggrund af korpora
2. Teori og praksis for hensigtsmæssig opbygning og udnyttelse af korpora

Hvad er et korpus?

Hvad er et korpus?

- ▶ „In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.“ Kennedy 1998

Hvad er et korpus?

- ▶ „In the language sciences a corpus is a body of **written text or transcribed speech** which can serve as a basis for linguistic analysis and description.“ Kennedy 1998

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a body of **written text or transcribed speech** which can serve as a basis for linguistic analysis and description.“ Kennedy 1998

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.“ Kennedy 1998

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.“ Kennedy 1998

almensprog
el. særsprog

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a **body of written text or transcribed speech** which can serve as a basis for **linguistic analysis and description.**“ Kennedy 1998

almensprog
el. særsprog

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a **body of written text or transcribed speech** which can serve as a basis for **linguistic analysis and description.**“ Kennedy 1998

almensprog
el. særsprog

stort
og balanceret

Hvad er et korpus?

i
digital form

- ▶ „In the language sciences a corpus is a **body of written text or transcribed speech** which can serve as a basis for **linguistic analysis and description.**“ Kennedy 1998

almensprog
~~el. særsprog~~

stort
og balanceret

Hvad er et korpus?

i
digital form

Referencekorpus:
en antaget repræsentativ
stikprøve af sproget

almensprog
~~el. særsprog~~

stort
og balanceret

Referencekorpora

Korpus	Tekster	Ord
DDO	43.000	40 mio.
Korpus 90	–	28 mio.
Korpus 2000	–	28 mio.

Referencekorpora

Korpus	Tekster	Ord
DDO	43.000	40 mio.
Korpus 90	–	28 mio.
Korpus 2000	–	28 mio.

Referencekorpora

Korpus	Tekster	Ord
DDO	43.000	40 mio.
Korpus 90	–	28 mio.
Korpus 2000	–	28 mio.

Referencekorpora

Korpus	Tekster	Ord
DDO	43.000	40 mio.
Korpus 90	–	28 mio.
Korpus 2000	–	28 mio.

Korpussammensætning
eksemplificeres ved DDO's korpus

Program

1. Tekstkorpora og referencekorpora

2. Korpussammensætning

3. Korpusopmærkning

4. Korpusundersøgelser

5. Fremtiden

Korpussammensætning

Tekstinfo

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Korpussammensætning

Tekstinfo

Domæne

Genre

66 hhv. 12

forskellige værdier, fx
geografi, musik, filosofi

131 hhv. 17

forskellige værdier, fx *roman,*
interview, essay

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Genre

131 hhv. 17
forskellige værdier, fx *roman,*
interw, essay

Medium

forskellige værdier, fx *bog, avis,*
dagbog

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Genre

131 hhv. 17
forskellige værdier, fx *roman,*
interview, essay

Medium

forskellige værdier, fx *bog, avis,*
almensprog
eller *fagligt sprog*

Sprogtype

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Genre

131 hhv. 17
forskellige værdier, fx *roman,*
interview, essay

Medium

forskellige værdier, fx *bog, avis,*
almensprog
eller *fagligt sprog*

Sprogtype

talesprog
eller *skriftsprog*

Udtryk

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Genre

131 hhv. 17
forskellige værdier, fx *roman,*
interview, essay

Medium

13
forskellige værdier, fx *bog, avis,*
almensprog
eller *fagligt sprog*

Sprogtype

talesprog
eller *skriftsprog*

reception
eller *produktion*

Udtryk

Aspekt

Korpussammensætning

Tekstinfo

Domæne

66 hhv. 12
forskellige værdier, fx
geografi, musik, filosofi

Genre

131 hhv. 17
forskellige værdier, fx roman,
interview, essay

Medium

13
forskellige værdier, fx bog, avis,
almensprog
eller fagligt sprog

Sprogtype

talesprog
eller skriftsprog
reception
eller produktion

Udtryk

Aspekt

Produktionsår

1983–
1992

Korpussammensætning

Sprogbrugerinfo

Korpussammensætning

Sprogbrugerinfo

Køn

*mand,
kvinde, ukendt*

Korpussammensætning

Sprogbrugerinfo

Køn

*mand,
kvinde, ukendt*

Fødselsår

*1880–
1990*

Korpussammensætning

Sprogbrugerinfo

Køn

*mand,
kvinde, ukendt*

Fødselsår

*1880–
1990*

Fødested

*frit
stednavn*

Korpussammensætning

Sprogbrugerinfo

Køn

*mand,
kvinde, ukendt*

Fødselsår

*1880–
1990*

Fødested

*frit
stednavn*

Dialektområde

11 regioner

Korpussammensætning

Sprogbrugerinfo

Køn

mand,
kvinde, ukendt

Fødselsår

1880–
1990

Fødested

frit
stednavn

Dialektområde

11 regioner

Udannelse

fri
betegnelse

Korpussammensætning

Sprogbrugerinfo

Køn

mand,
kvinde, ukendt

Fødselsår

1880–
1990

Fødested

frit
stednavn

Dialektområde

11 regioner

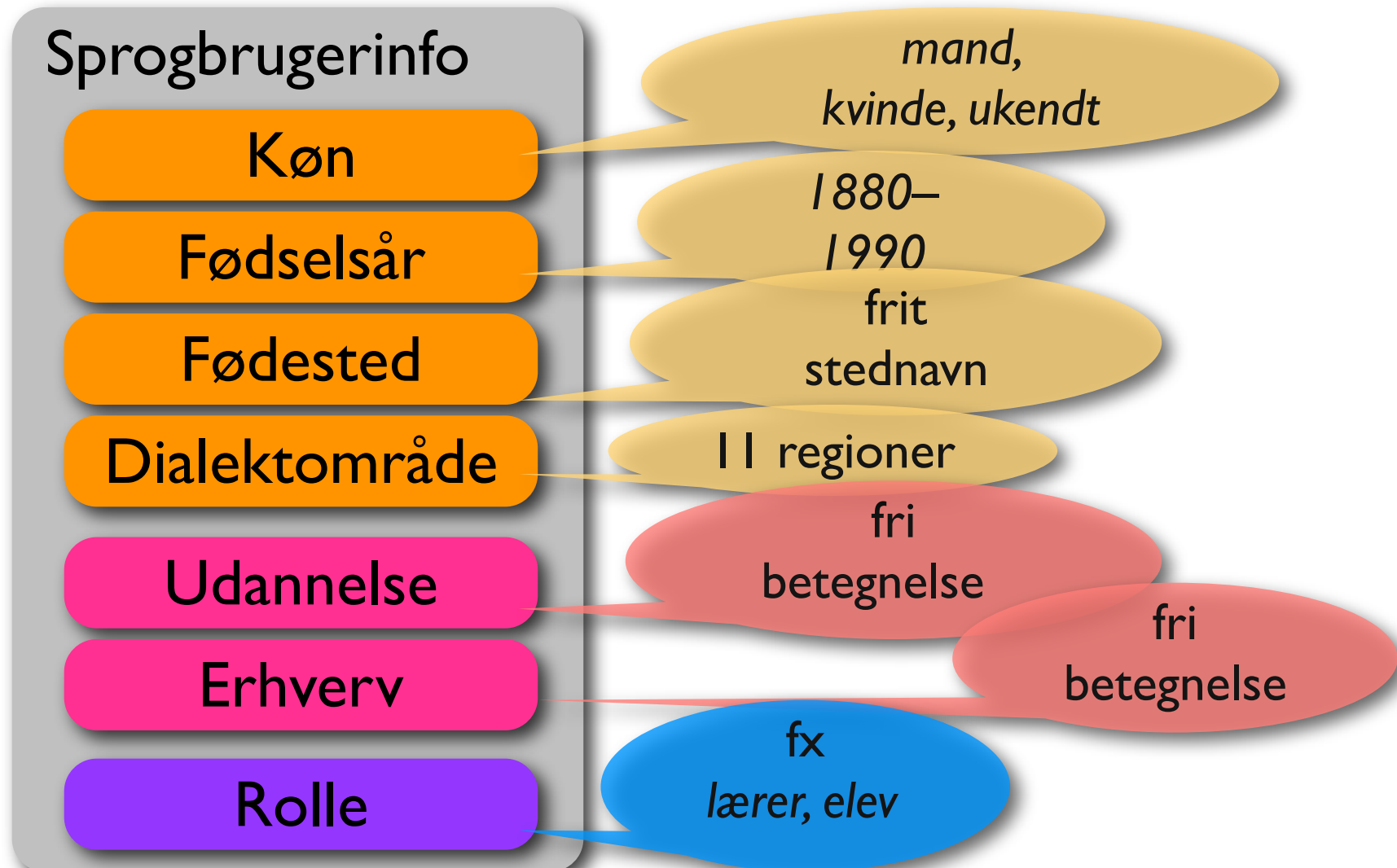
Udannelse

fri
betegnelse

Erhverv

fri
betegnelse

Korpussammensætning



Korpussammensætning

Sprogbrugerinfo

Køn

Fødselsår

Fødested

Dialektområde

Udannelse

Erhverv

Rolle

Korpussammensætning

Header

Tekstinfo

Domæne

Genre

Medium

Sprogtype

Udtryk

Aspekt

Produktionsår

Sprogbrugerinfo

Køn

Fødselsår

Fødested

Dialektområde

Udannelse

Erhverv

Rolle

Korpussammensætning

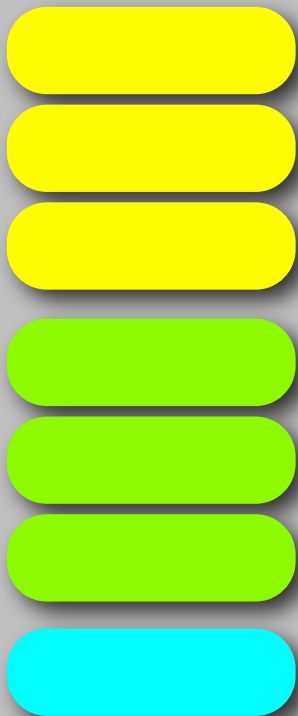


Korpussammensætning

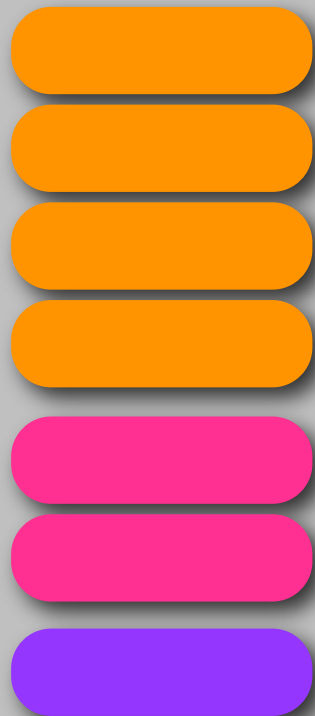
Korpusenhed

Header

Tekstinfo



Sprogbruger



Tekst

```
<p><f>NY DUFT.</f> Den er sødlig.  
Eksotisk. Så forførende, at den lokker  
til romantisk eventyr.</p>  
<p>Gracious! Din nye Impulse. Med  
duften, som er in lige nu hos de fineste  
parfumehuse i verden.</p>  
<p>Og den milde deovirkning, der holder  
dig frisk og dejlig.</p>  
<p>Hele dagen.</p>  
<p><f>GRACIOUS.</f></p>  
<p>Deodorant og parfume. Altid en duft  
for dig.</p>  
<p><f>impulse</f></p>  
<p>perfumed deodorant</p>  
<p>Gracious</p>
```

Korpussammensætning

```
<Korpusenhed>
<Header>
<TxtOpl>
<Id>RTng</Id><Restr><Ano>-</Ano><DD0>-</DD0></Restr><Ttit>-</Ttit><Vtit>Vi Unge</
Vtit><Forl>Specialbladsforlaget</Forl><Dat><Dg>-</Dg><Md>3</Md><År>88</År><Si>-</Si></Dat><Lo>3:</
Lo><AlFa>a</AlFa><SkTa>s</SkTa><RePr>r</RePr><Arel>vu</Arel><Medi>bl</Medi>
<Genr>rekl</Genr><GnTy>ann</GnTy><Emne>65</Emne><Grp>ViUnge-rekl1KK</Grp><Num>1</
Num><Fil>VIUNREKL</Fil><Omf>715</Omf>
</TxtOpl>
<SpbOpl>
<EfN>?</EfN><FoN>?</FoN><Køn>?</Køn><FøÅr><År>?</År><Si>-</Si></FøÅr><FøS>?</FøS><Bop>?</
Bop><Reg>?</Reg><Udd>?</Udd><Erh>?</Erh><SpV>i</SpV><Rol>?</Rol>
</SpbOpl>
</Header>

<Tekst ID=RTng>
<p><f>DU HAR ALDRIG SET HAM FØR</f></p><p><f>PLUDELIG GI'R HAN DIG BLOMSTER</f></
p><p><f>impulse</f></p><p><f>NY DUFT.</f> Den er sødlig. Eksotisk. Så forførende, at den lokker
til romantisk eventyr.</p><p>Gracious! Din nye Impulse. Med duften, som er in lige nu hos de
fineste parfumehuse i verden.</p><p>Og den milde deovirkning, der holder dig frisk og dejlig.</p>
<p>Hele dagen.</p><p><f>GRACIOUS.</f></p><p>Deodorant og parfume. Altid en duft for dig.</
p><p><f>impulse</f></p><p>perfumed deodorant</p><p>Gracious</p>
</Tekst>
</Korpusenhed>
```

Korpussammensætning

```
<Korpusenhed>
<Header>
<TxtOpl>
<Id>RTng</Id><Restr><Ano>-</Ano><DDO>-</DDO></Restr><Ttit>-</Ttit><Vtit>Vi Unge</
Vtit><Forl>Specialbladsforlaget</Forl><Dat><Dg>-</Dg><Md>3</Md><År>88</År><Si>-</Si></Dat><Lo>3:</
Lo><AlFa>a</AlFa><SkTa>s</SkTa><RePr>r</RePr><Arel>vu</Arel><Medi>bl</Medi>
<Genr>rekl</Genr><GnTy>ann</GnTy><Emne>65</Emne><Grp>ViUnge-rekl1KK</Grp><Num>1</
Num><Fil>VIUNREKL</Fil><Omf>715</Omf>
</TxtOpl>
<SpbOpl>
<EfN>?</EfN><FoN>?</FoN><Køn>?</Køn><FøÅr><År>?</År><Si>-</Si></FøÅr><FøS>?</FøS><Bop>?</
Bop><Reg>?</Reg><Udd>?</Udd><Erh>?</Erh><SpV>i</SpV><Rol>?</Rol>
</SpbOpl>
</Header>
<Tekst ID=RTng>
<p><f>DU HAR ALDRIG SET HAM
p><p><f>impulse</f></p><p><f>
til romantisk eventyr.</p><p><f>
fineste parfumehuse i verden
<p>Hele dagen.</p><p><f>GRAC
p><p><f>impulse</f></p><p>pe
</Tekst>
</Korpusenhed>
```

Tekstinfo

Domæne

Genre

Medium

Sprogtype

Udtryk

Aspekt

Produktionsår

Sprogbrugerinfo

Køn

Fødselsår

Fødested

Dialektområde

Udannelse

Erhverv

Rolle

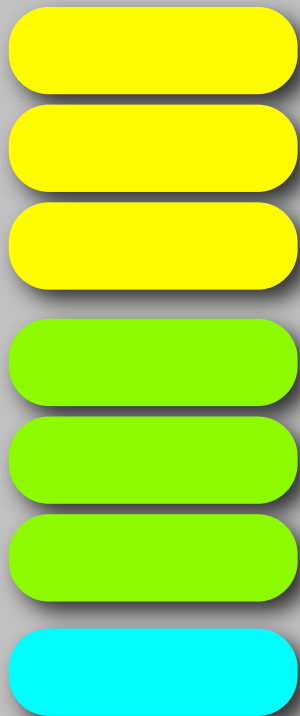
```
DIG BLOMSTER</f></
k. Så forførende, at den lokker
en, som er in lige nu hos de
r holder dig frisk og dejlig.</p>
. Altid en duft for dig.</
```

Korpussammensætning

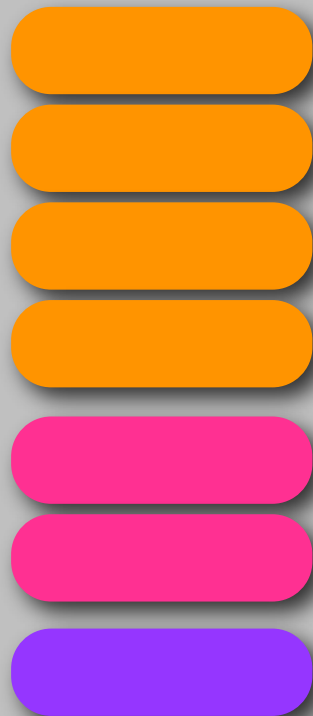
Korpusenhed

Header

Tekstinfo



Sprogbruger



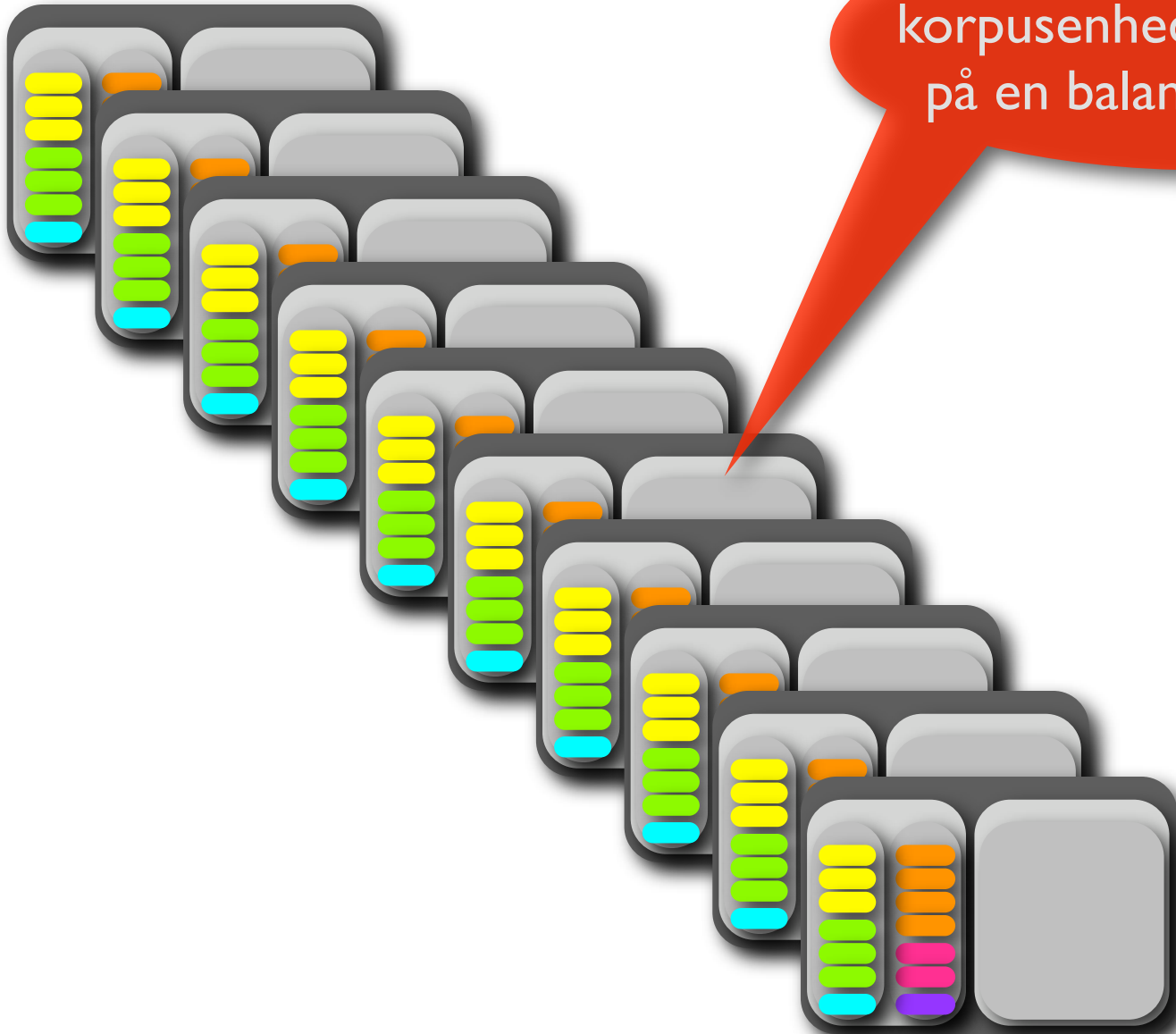
Tekst

```
<p><f>NY DUFT.</f> Den er sødlig.  
Eksotisk. Så forførende, at den lokker  
til romantisk eventyr.</p>  
<p>Gracious! Din nye Impulse. Med  
duften, som er in lige nu hos de fineste  
parfumehuse i verden.</p>  
<p>Og den milde deovirkning, der holder  
dig frisk og dejlig.</p>  
<p>Hele dagen.</p>  
<p><f>GRACIOUS.</f></p>  
<p>Deodorant og parfume. Altid en duft  
for dig.</p>  
<p><f>impulse</f></p>  
<p>perfumed deodorant</p>  
<p>Gracious</p>
```

Korpussammensætning

43.000

korpusenheder sammensat
på en balanceret måde

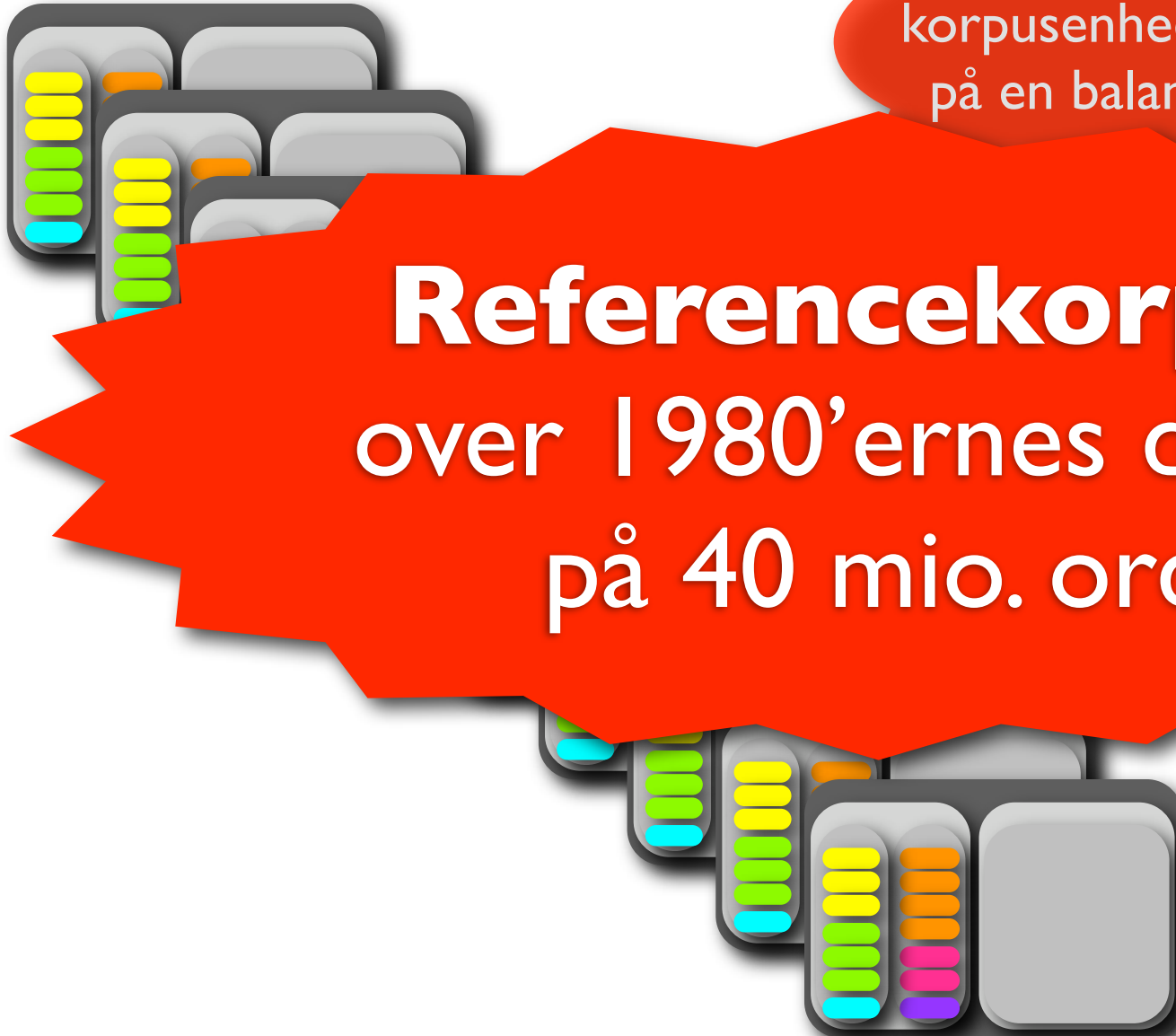


Korpussammensætning

43.000

korpusenheder sammensat
på en balanceret måde

Referencekorpus
over 1980'ernes dansk
på 40 mio. ord



Program

1. Tekstkorpora og referencekorpora
2. Korpussammensætning
3. Korpusopmærkning
4. Korpusundersøgelser
5. Fremtiden

Korpusopmærkning

- På tekstniveau
- På ordniveau
- På andre niveauer: sætning, morfem etc.

Korpusopmærkning

A red speech bubble with a white shadow, pointing towards the first bullet point. It contains the text "Behandlet ifm. med headerne".

Behandlet
ifm. med headerne

- På tekstniveau
- På ordniveau
- På andre niveauer: sætning, morfem etc.

Korpusopmærkning

Behandlet
ifm. med headerne

- På tekstniveau
- På ordniveau
- På andre niveauer: sætning, morfem etc.

Kommer
vi ikke ind på

Korpusopmærkning

- På tekstniveau
- På ordniveau
- På andre niveauer: sætning, morfem etc.

Behandlet
ifm. med headerne

Eksemplificeres
ved Korpus 2000

Kommer
vi ikke ind på

Abstraktionsniveauer

- Udgangspunkt: løbende tekst
 - ▶ Token- og sætningsopdeling
 - ▶ Lemmatisering
 - ▶ Ordklassetagging
 - ▶ Syntaktisk parsning
 - ▶ Semantisk opmærkning

Abstraktionsniveauer

- Udgangspunkt: løbende tekst

▶ Token- og sætningsopdeling

▶ Lemmatisering

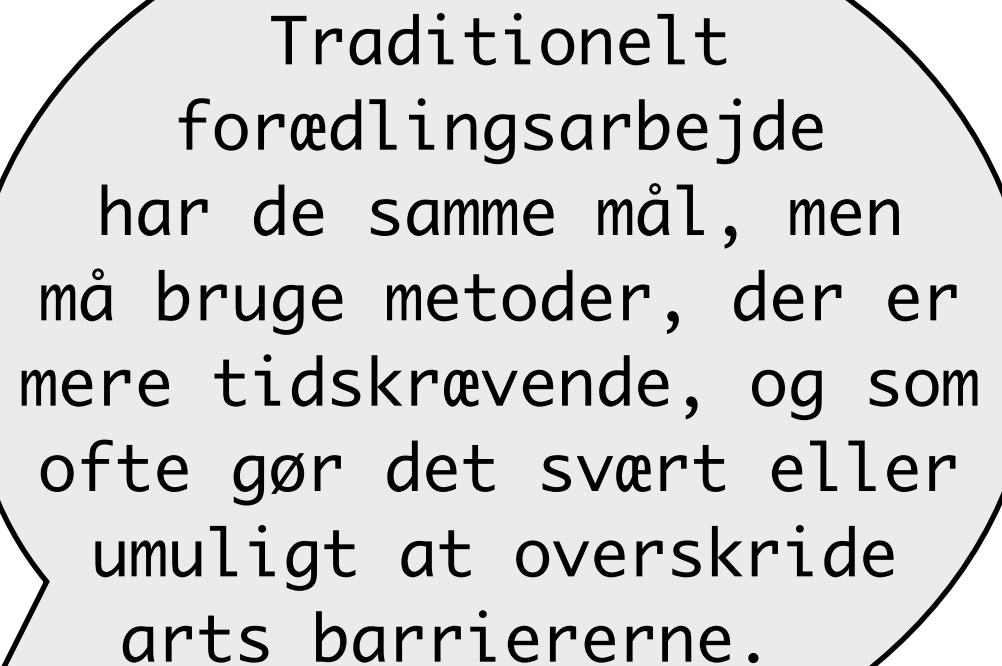
▶ Ordklassetagging

▶ Syntaktisk parsning

▶ Semantisk opmærkning

Disse 3 niveauer
ser vi nærmere på

Tokenopdeling



Traditionelt
forædlingsarbejde
har de samme mål, men
må bruge metoder, der er
mere tidskrævende, og som
ofte gør det svært eller
umuligt at overskride
arts barriererne.

Tokenopdeling

Traditionelt
forædlingsarbejde
har de samme mål, men
må bruge metoder, der er
mere tidskrævende, og som
ofte gør det svært eller
umuligt at overskride
arts barriererne.

Tokenopdeling

Traditionelt
forædlingsarbejde
har de samme mål, men
må bruge metoder, der er
mere tidskrævende, og som
ofte gør det svært eller
umuligt at overskride
arts barriererne.

Tokenopdeling

Traditionelt	-
forædlingsarbejde	-
har	-
de	-
samme	-
mål	,-
men	-
må	-
bruge	-
metoder	,-
der	-
er	-
mere	-
tidskrævende	,-

Traditionelt_
forædlingsarbejde_
har_ de_ samme_ mål, men_
må_ bruge_ metoder, der_ er_
mere_ tidskrævende, og_ som_
ofte_ gør_ det_ svært_ eller_
umuligt_ at_ overskride_
arts_ barriererne.

Tokens

Tokenadskillere

Lemmatisering

Traditionelt	—
forædlingsarbejde	—
har	—
de	—
samme	—
mål	,—
men	—
må	—
bruge	—
metoder	,—
der	—
er	—
mere	—
tidskrævende	,—

Lemmatisering

Traditionelt	–	traditionel
forædlingsarbejde	–	forædlingsarbejde
har	–	have
de	–	den
samme	–	samme
mål	,–	mål
men	–	men
må	–	måtte
bruge	–	bruge
metoder	,–	metode
der	–	der
er	–	være
mere	–	meget
tidskrævende	,–	tidkrævende



Lemmaformer
(grundformer)

Lemmatisering

Traditionelt	–	traditionel
forædlingsarbejde	–	forædlingsarbejde
har	–	have
de	–	den
samme	–	samme
mål	,–	mål
men	–	men
må	–	måtte
bruge	–	bruge
metoder	,–	metode
der	–	der
er	–	være
mere	–	meget
tidskrævende	,–	tidskrævende

Lemmasformer
(grundformer)

Lemmatisering

Traditionelt	–	traditionel
forædlingsarbejde	–	forædlingsarbejde
har	–	have
de	–	den
samme	–	samme
mål	, –	mål
men	–	men
må	–	måtte
bruge	–	bruge
metoder	, –	metode
der	–	der
er	–	være
mere	–	meget
tidskrævende	, –	tidkrævende

Lemmasformer
(grundformer)

Lemmatisering
forudsætter et
fuldformsleksikon og en
disambigueringsrutine

Ordklassetagging

Traditionelt	–	traditionel
forædlingsarbejde	–	forædlingsarbejde
har	–	have
de	–	den
samme	–	samme
mål	,–	mål
men	–	men
må	–	måtte
bruge	–	bruge
metoder	,–	metode
der	–	der
er	–	være
mere	–	meget
tidskrævende	,–	tidskrævende

Ordklassetagging

Traditionelt	-	traditionel	ADJ	NEU S IDF NOM
forædlingsarbejde	-	forædlingsarbejde	N	NEU S IDF NOM
har	-	har	V	PR AKT
de	-	de	ART	nG P DEF
de samme	-	de samme	DET	nG nN NOM
mål	-	mål	N	NEU P IDF NOM
men	-	men	KC	
må	-	måtte	V	PR AKT
bruge	-	bruge	V	INF AKT
metoder	, -	metode	N	UTR P IDF NOM
der	-	der	INDP	nG nN NOM
er	-	være	V	PR AKT
mere	-	meget	ADV	COM
tidskrævende	, -	tidskrævende	ADJ	nG nN nD NOM

Ordklassetagging
forudsætter et
fuldformsleksikon og en
disambigueringsrutine

Ordklassetags

Bøjningstags

Tekstformat

Traditionelt	–	traditionel	ADJ	NEU S IDF NOM
forædlingsarbejde	–	forædlingsarbejde	N	NEU S IDF NOM
har	–	have	V	PR AKT
de	–	den	ART	nG P DEF
samme	–	samme	DET	nG nN NOM
mål	,–	mål	N	NEU P IDF NOM
men	–	men	KC	
må	–	måtte	V	PR AKT
bruge	–	bruge	V	INF AKT
metoder	,–	metode	N	UTR P IDF NOM
der	–	der	INDP	nG nN NOM
er	–	være	V	PR AKT
mere	–	meget	ADV	COM
tidskrævende	,–	tidskrævende	ADJ	nG nN nD NOM

Tekstformat

Traditionelt	-	traditionel	ADJ	NEU S IDF NOM
forædlingsarbejde	-	forædlingsarbejde	N	NEU S IDF NOM
har	-	have	V	PR AKT
de	-	den	ART	nG P DEF
samme	-	samme	DET	nG nN NOM
mål	, -	mål	N	NEU P IDF NOM
me	-	men	KC	
må	-	måtte	V	PR AKT
bruge	-	bruge	V	INF AKT
metoder	, -	metode	N	UTR P IDF NOM
der	-	der	INDP	nG nN NOM
er	-	være	V	PR AKT
mere	-	meget	ADV	COM
tidskrævende	, -	tidskrævende	ADJ	nG nN nD NOM

Tokens

Tekstformat

Traditionelt
forædlingsarbejde
har
de
samme
mål
me
må
bruge
metoder
der
er
mere
tidskrævende

Tokens

traditionel	ADJ	NEU S IDF NOM
forædlingsarbejde	N	NEU S IDF NOM
have	V	PR AKT
den	ART	nG P DEF
samme	DET	nG nN NOM
, mål	N	NEU P IDF NOM
men		
måtte		KT
bruge	V	INF AKT
, metode	N	UTR P IDF NOM
der	INDP	nG nN NOM
være	V	PR AKT
meget	ADV	COM
, tidskrævende	ADJ	nG nN nD NOM

Attributter

Program

1. Tekstkorpora og referencekorpora
2. Korpussammensætning
3. Korpusopmærkning
4. Korpusundersøgelser
5. Fremtiden

Søgning i korpus

- En hvilken som helst kombination af tokens og tokenattributter
- Headeroplysninger kan inddrages

Søgning i Korpus 2000

- En hvilken som helst kombination af tokens og tokenattributter
- Headeroplysninger kan inddrages

Søgning i Korpus 2000

Visse
begrænsninger pga.
„brugervenlighed“

- En hvilken som helst kombination af tokens og tokenattributter
- Headeroplysninger kan inddrages

Ikke
muligt

Hvad er Korpus 2000?

- Referencekorpus over dansk sprog omkring år 2000
- Omfang på 28 mio. tokens
- Sammenlignende undersøgelser med DDO's korpus (Korpus 90)

Hvad er Korpus 2000?



The image shows a screenshot of a web browser window displaying the Korpus 2000 website. The browser's address bar shows the URL <http://korpus.dsl.dk/korpus2000/indgang.php>. The website has a dark blue header with the DSL logo (three lions) and the text "Korpus 2000 er udarbejdet af Det Danske Sprog- og Litteraturselskab DSL". A navigation menu includes "Opslag", "DSL", "Lingvistik", "Konferencer", "Info", and "Mac".

The main content area features a large white box with the heading "In English" and the text "Se hvordan vi bruger sproget - slå op i Korpus 2000!". Below this is a search input field labeled "Indtast ord eller vending:" and a button "Slå op i korpus". A "Hjælp" link is also visible.

A large blue banner is overlaid on the bottom half of the page, containing the text www.korpus2000.dk.

At the bottom, there is a section titled "En del af Ordnet.dk" with a paragraph of text: "Ordnet.dk er et projekt der i løbet af en seksårig periode (2004-2010) vil udvikle et sprogligt værktøj der knytter flere af DSL's ordbøger sammen og forbinder dem med tekstkorpora så der skabes helt nye muligheder for opslag og effektive søgninger i et meget stort materiale. En af opgaverne er en web-version af [Ordbog over det danske Sprog](#)."

Søgning på lemma

Se hvordan vi bruger sproget
- slå op i Korpus 2000!

Indtast ord eller vending:

regn

Slå op i korpus

[Hjælp](#)

Søgning på lemma

Gå
ind på
www.korpus2000.dk
og indtast et ord

Se hvordan vi bruger sproget
- slå op i Korpus 2000!

Indtast ord eller vending:

regn

Slå op i korpus

Hjælp

Klik
her

Søgning på lemma



Vis hyppigheder, eksempler m.m. for ...


regn 😊, sb. - med alle tilhørende bøjningsformer

regne 😊, vb. - med alle tilhørende bøjningsformer

regn - kun den indtastede form

Søgning på lemma

Det
indtastede regn kan både være
en form af regn, sb. eller regne, vb.
Vælg ønsket lemma...

 Vælg ord


Vis hyppigheder, eksempler m.m. for ...

- regn** 😊, sb. - med alle tilhørende bøjningsformer
- regne** 😊, vb. - med alle tilhørende bøjningsformer
- regn** - kun den indtastede form











Klik
her

Søgning på lemma

← Vælg ord Hyppighed Nabo-ord Vendinger Orddele Stavning

Klik på  i oversigten for at se eksemplerne fra korpus for pågældende form.

Hyppighed i korpus regn 😊, sb.

	Korpus 2000	Korpus 90
regn	●●●●●●●● 	●●●●●●●● 
regns	●●●●●●●●	 ●●●●●●●● 
regnen	●●●●●●●● 	●●●●●●●● 
regnets	●●●●●●●● 	●●●●●●●● 
regn 😊, sb. alle former	●●●●●●●● 	●●●●●●●● 

Søgning på lemma

Klik på i oversigten for at se eksemplerne fra korpus gældende form.

Hyppighed i korpus regn her for at se en konkordans over formen *regnen*, sb. i Korpus 2000

	Korpus 2000	Korpus 90
regn	●●●●●●	●●●●●●
regns	●●●●●●	●●●●●●
regnen	●●●●●●	●●●●●●
regnens	●●●●●●	●●●●●●
regn , sb. alle former	●●●●●●	●●●●●●

Klik her for at se alle former af lemmaet *regn*, sb.

Søgning på lemma

op midt på gaden , og fyrede en **regn** af bønner afsted mod et par sirener ,
ind_i døren , som flyver op i en **regn** af gnister . Dørens fløje rammer væg
idseligt vindpust førte med ét en **regn** af gnister over_mod nabohuset , og li
i en syvårs dreng sender han en **regn** af kugler ud i kontoret . " Fjenden
k efter brasede det sammen i en **regn** af gnister ligesom retshusets kuppel l
r dog for intet at regne mod den **regn** af roser , der er væltet ned over
nhavn var sidste år udsat for en **regn** af klager . Også landsformanden for d
konkret roder sig ud i - under en **regn** af vrede læggekartofler ! Mikkel_Hede
ag til finansloven i_dag . En mild **regn** af euro daler i_dag ned over det dans
det tyste og lavloftede rum . En **regn** af gnister lyner gennem mørket , og
ti og brandvæsen mødes med en **regn** af sten , når de rykker ud på
es . Lige_så kontroversiel er den **regn** af gaver , som Clinton-parret sikrede
ke film . Og at dømme efter den **regn** af priser , som Peter_Schönau_Fog h
s_Angeles til s. Han forventer en **regn** af sagsanlæg fra firmaer , som er blev
uplen med klokken sammen i en **regn** af flammer og gnister . Hans tunge kl
 , og han blev modtaget med en **regn** af spørgsmål om hans kontrovers me
 , andet end at han tog imod en **regn** af slag . Og dermed er der risiko
i deres alder , levede vi i en **regn** af kugler . Vi vidste ikke , hvad
der i to timer blev udsat for en **regn** af skudsalver fra et elitepolitikorps . I
å . De fleste dage bliver grå med **regn** af_og_til . Temperaturer mellem nul

Søgning på lemma

Resultatet
er en KWIC-
konkordans,
sorteret

op midt på gaden , og fyrede en **regn** af bønner afsted mod et par sirener ,
ed_i døren , som flyver op i en **regn** af gnister . Dørens fløje rammer væg
idseligt vindpust førte med ét en **regn** af gnister over_mod nabohuset , og li
en syvårs dreng sender han en **regn** af kugler ud i kontoret . " Fjenden
efter forasede det sammen i en **regn** af gnister ligesom retshusets kuppel l
r dog for intet at regne mod en **regn** af roser , der er væltet ned over
nhavn var sidste år udsat for en **regn** af klager . Også landsformanden for d
konkret roder sig ud i - under en **regn** af vrede læggekartofler ! Mikkel_Hede
ag til finansloven i_dag . En mild **regn** af euro daler i_dag ned over det dans
det tyste og lavloftede rum . En **regn** af gnister lyner gennem mørket , og
ti og brandvæsen mødes med en **regn** af sten , når de rykker ud på
es . Lige_så kontroversiel er den **regn** af gaver , som Clinton-parret sikrede
ke film . Og at dømme efter den **regn** af priser som Peter_Schönau_Fog h
s_Angeles til s. Han forventer en **regn** af sagsanlæg fra firmaer , som er
uplen med klokken sammen i en **regn** af flammer og gnister . Hans tunge k
 , og han blev modtaget med en **regn** af spørgsmål om hans kontroversie
 , andet end at han tog imod en **regn** af slag . Og dermed er der risiko
 i deres alder , levede vi i en **regn** af kugler . Vi vidste ikke , hvad
der i to timer blev udsat for en **regn** af skudsaver fra et elitepolitikorps . I
å . De fleste dage bliver grå med **regn** af_og_til . Temperaturer mellem nul

KWIC
= keyword in context

Sætningskløvning

Se hvordan vi bruger sproget
- slå op i Korpus 2000!

Indtast ord eller vending:

det er N som|der V

Slå op i korpus

[Hjælp](#)

Sætningskløvning

Indtast
en gruppe af ord. *N* og *V*
er pladsholdere. Den lodrette
streg betyder *eller*.

Se hvordan vi bruger sproget
- slå op i Korpus 2000!

Indtast ord eller vending:

det er N som|der V

Slå op i korpus

[hjælp](#)

Klik
her

Sætningskløvning

Rediger ordgruppe

Vis eksemplerne fra Korpus 2000

med følgende valgte ord eller ordklasser i den viste rækkefølge

'det' være, vb. sb. 'som|der' vb.

Sætningskløvning

Klik
her

Rediger ordgruppe

Vis eksemplerne fra Korpus 2000

med følgende valgte ord eller ordklasser i den viste rækkefølge

'det' være, vb. sb. 'som|der' vb.

Ret til
vha. rullemenuerne:
er → være, vb.
N → sb.
V → vb.

Sætningskløvning

ernes slægt , **det** er celloen der brummer , når han spiller
ildnes , fordi **det** var Blackman der angreb først , men landsretten
ille være , at **det** var Byrådet der anviste , hvad , hvor
tvivl om , at **det** var sejren der gled favoritten Cowboybuks_Frøkj
profil for , at **det** er aktionæerne der skal have gevinster og styre
tering , men **det** er målet der gør midlet til udtryk .
ippen , men **det** er bankerne der bestemmer , hvornår den skal
dreorloven , **det** er kvinderne der får den laveste løn .
ation , hvor **det** var politiet som løb , og ikke mig
skatten , og **det** er staten der betaler moderniseringen . Alarmcer
r hedder , at **det** er øjeblikket der tæller , hvilket hænger godt
is ord , men **det** er ord som har skabt voldsom frygt hos
omsorg . Og **det** er kærligheden der er adgangen til opstandelsen
 . Men , skal **det** være samfundet som betaler for denne service .
 : revisor , og **det** er bestyrelsen som har ansvaret . Jeg var
 ikke på , at **det** er pengene der afholder folk i Danmark fra
 . Og selvom **det** er sejrherren der skriver historien , så er
 r stor , men **det** er tøjet der er for småt . Hun

Sætningskløvning

ernes slægt , **det** er celloen der brummer , når han spiller
ildnes , fordi **det** var Blackman der angreb først , men landsretten
ille være , at **det** var Byrådet der anviste , hvad , hvor
tvivl om , at **det** var sejren der gled favoritten Cowboybuks_Frøkj
profil for , at **det** er aktionæerne der skal have gevinster og styre
rering , men **det** er målet der gør midlet til udtryk .
ippen , men **det** er bankerne der bestemmer , hvornår den skal
dreorloven , **det** er kvinderne der får den laveste løn .
ation , hvor **det** var politiet som løb , og ikke mig
skatten , og **det** er staten der betaler moderniseringen . Alarmce
r hedder , at **det** er øjeblikket der tæller , hvilket hænger godt
is ord , men **det** er ord som har skabt voldsom frygt hos
omsorg . Og **det** er kærligheden der er adgangen til opstandelsen
Men , skal **det** være samfundet som betaler for denne service
revisor , og **det** er bestyrelsen som har ansvaret . Jeg var
ikke på , at **det** er pengene der afholder folk i Danmark fra
 . Og selvom **det** er sejrherren der skriver historien , så er
r stor , men **det** er tøjet der er for små . Hun

OBS!
Vær kritisk over
for resultatet!

Resulterende
KWIC-konkordans

Mere om søgning

- På www.korpus2000.dk kan man også
 - ▶ Søge med regulære udtryk
 - ▶ Se ordlister
 - ▶ Lave kollokationsundersøgelser
- Se mere på korpus.dsl.dk/staff/ja/papers/prag2006/pres_uniprag.pdf

Program

1. Tekstkorpora og referencekorpora
2. Korpussammensætning
3. Korpusopmærkning
4. Korpusundersøgelser
5. Fremtiden

Hvad er ordnet.dk?

The image shows a screenshot of a web browser window displaying the homepage of the Korpus 2000 project. The browser's address bar shows the URL <http://korpus.dsl.dk/korpus2000/indgang.php>. The page features a dark blue header with the 'DSL' logo and the text 'Korpus 2000 er udarbejdet af Det Danske Sprog- og Litteraturselskab DSL'. A navigation menu on the left includes links for 'Forside', 'Om Korpus 2000', 'Sproginspiration', 'Diskussionsforum', 'Download', 'Hjælp', 'Kontakt', 'Links', 'Nyhedsbrev', 'Summary', and 'Zusammenfassung'. The main content area has a white background and contains the heading 'In English' followed by 'Se hvordan vi bruger sproget - slå op i Korpus 2000!'. Below this is a search input field with the placeholder text 'Indtast ord eller vending:' and a 'Slå op i korpus' button. A 'Hjælp' link is also present. The text below explains that Korpus 2000 is a digital dictionary showing how Danish is used around the year 2000, based on a massive text collection of over 50 million words. It also mentions the 'Ordned.dk' project, which aims to create a linguistic tool by combining DSL's dictionaries.

Korpus 2000 er udarbejdet af Det Danske Sprog- og Litteraturselskab DSL

DSL

Korpus 2000

In English

Se hvordan vi bruger sproget - slå op i Korpus 2000!

Indtast ord eller vending:

Slå op i korpus

Hjælp

Hvad er Korpus 2000?

Et digitalt opslagsværk der viser hvordan dansk sprog faktisk bliver brugt - omkring år 2000 og omkring 1990. Opbygget omkring en kæmpe samling tekster med langt over 50 millioner ord. Den største og eneste samling af sin art i Danmark. Frit tilgængelig for alle! Læs [mere...](#)

En del af Ordned.dk

[Ordned.dk](#) er et projekt der i løbet af en seksårig periode (2004-2010) vil udvikle et sprogligt værktøj der knytter flere af DSL's ordbøger sammen og forbinder dem med tekstkorpora så der skabes helt nye muligheder for opslag og effektive søgninger i et meget stort materiale. En af opgaverne er en web-version af [Ordbog over det danske Sprog](#).

Hvad er ordnet.dk?

The screenshot shows a web browser window titled "Korpus 2000" with the URL <http://korpus.dsl.dk/korpus2000/indgang.php>. The page features a navigation menu on the left with items like "Forside", "Om Korpus 2000", "Sproginspiration", "Diskussionsforum", "Download", "Hjælp", "Kontakt", "Links", "Nyhedsbrev", "Summary", "Zusammenfassung", and "Printvenlig side". The main content area is titled "In English" and "Se hvordan vi bruger sproget - slå op i Korpus 2000!". It includes a search input field labeled "Indtast ord eller vending:" with a "Slå op i korpus" button and a "Hjælp" link. A yellow speech bubble points to the search area with the text "Igangværende DSL-projekt". Below the search area, there is a section titled "Hvad er Korpus 2000?" with a paragraph of text. At the bottom, a red-bordered box highlights a section titled "En del af Ordnet.dk" with a paragraph of text.

Korpus 2000

Korpus 2000 er udarbejdet af Det Danske Sprog- og Litteraturselskab DSL

In English

Se hvordan vi bruger sproget - slå op i Korpus 2000!

Indtast ord eller vending:

Slå op i korpus
Hjælp

Hvad er Korpus 2000?

Et digitalt opslagsværk der viser hvordan dansk sprog faktisk bliver brugt - omkring år 2000 og omkring 1990. Opbygget omkring en kæmpe samling tekster med langt over 50 millioner ord. Den største og eneste samling af sin art i Danmark. Frit tilgængelig for alle! Læs mere...

En del af Ordnet.dk

Ordnet.dk er et projekt der i løbet af en seksårig periode (2004-2010) vil udvikle et sprogligt værktøj der knytter flere af DSL's ordbøger sammen og forbinder dem med tekstkorpora så der skabes helt nye muligheder for opslag og effektive søgninger i et meget stort materiale. En af opgaverne er en web-version af Ordbog over det danske Sprog.

Hvad er ordnet.dk?



En del af Ordnet.dk

Ordnet.dk er et projekt der i løbet af en seksårig periode (2004-2010) vil udvikle et sprogligt værktøj der knytter flere af DSL's ordbøger sammen og forbinder dem med tekstkorpora så der skabes helt nye muligheder for opslag og effektive søgninger i et meget stort materiale. En af opgaverne er en web-version af **Ordbog over det danske Sprog**.

Fremtiden

- *ordnet.dk* etablerer en samlet tilgang til
 - ▶ Korpus 2000 og Korpus 90
 - ▶ Den Danske Ordbog
 - ▶ Ordbog over det danske Sprog
- Følg med på dsl.dk/ordboger-og-sprogteknologi/ordnet.dk