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...
www.dsl.dk

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/
Google
Opslag
DSL
Lingvistik
Konferencer
Info
Mac
Sideoversigt
Tilgængelighed
Nyhedsarkiv
Om DSL
Kontakt
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 delemner, 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
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

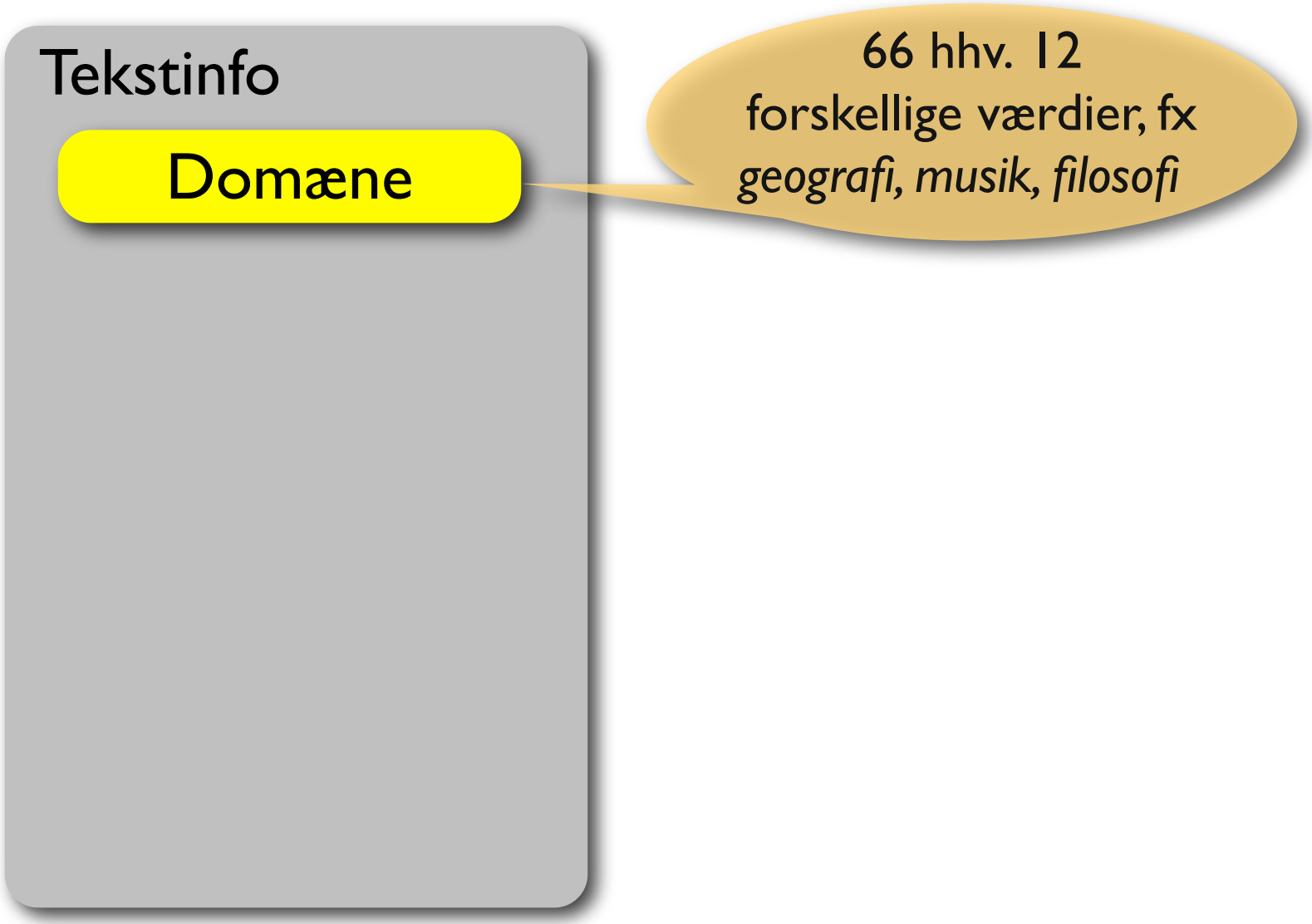
Program

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

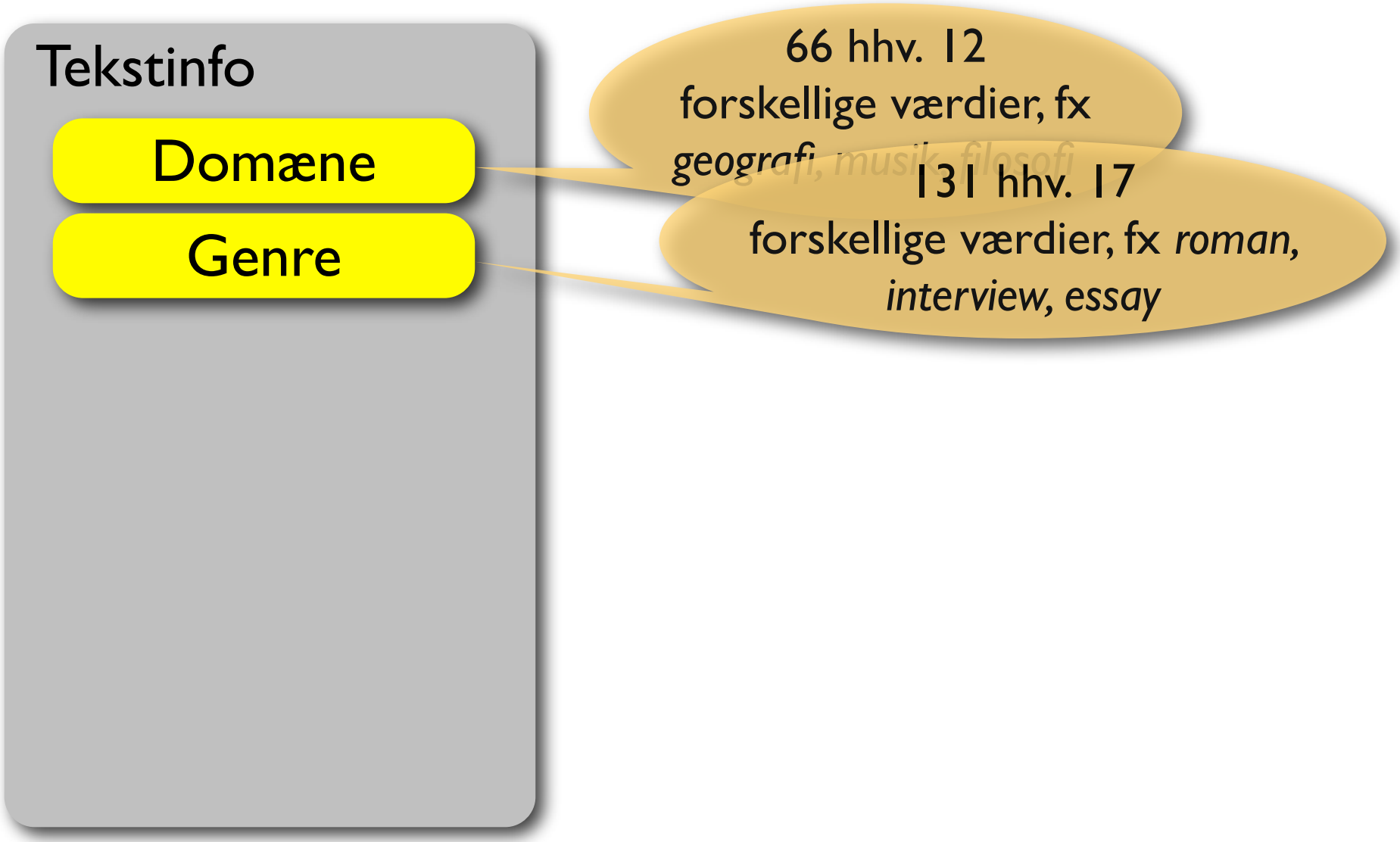
Korpussammensætning

Tekstinfo

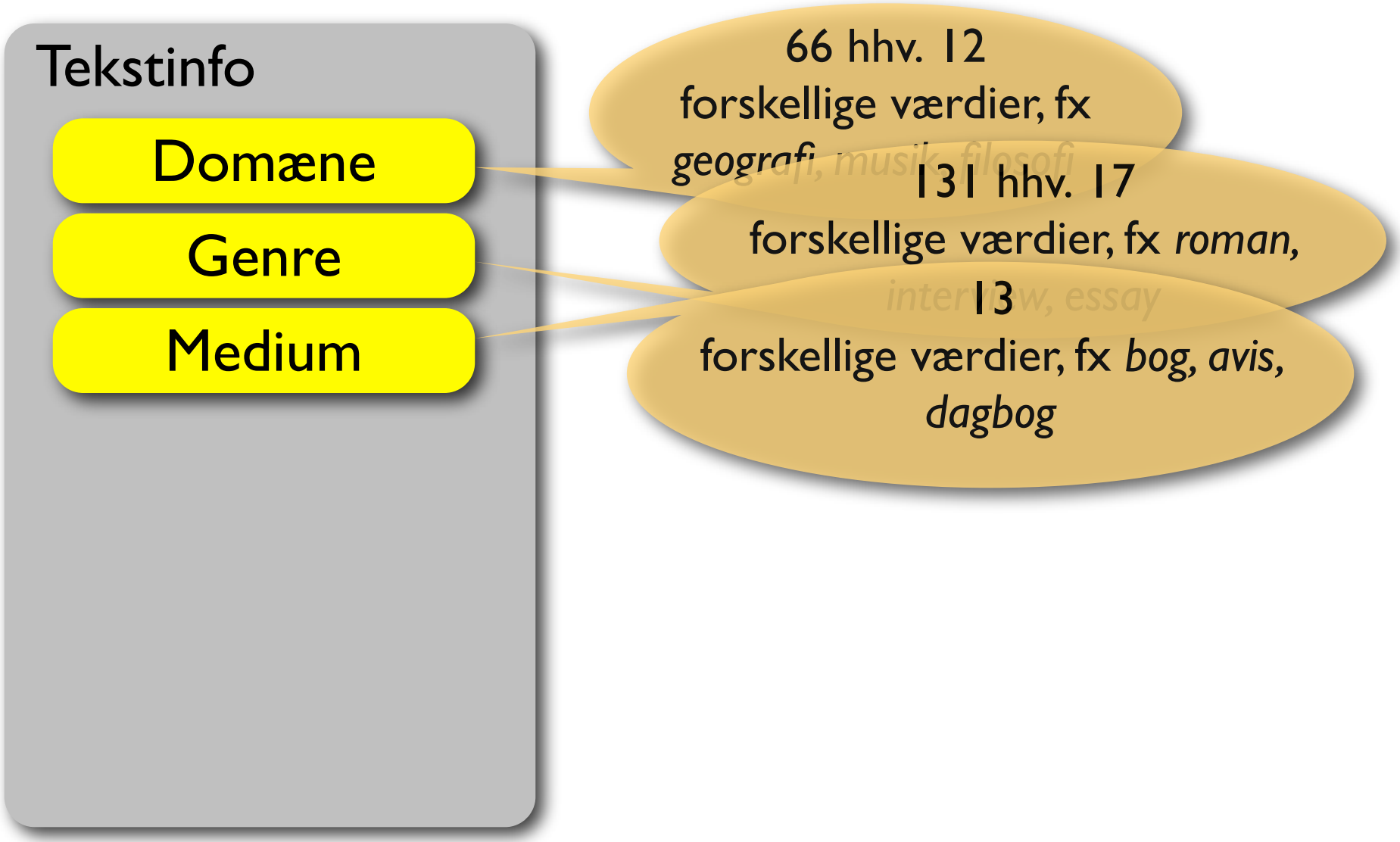
Korpussammensætning



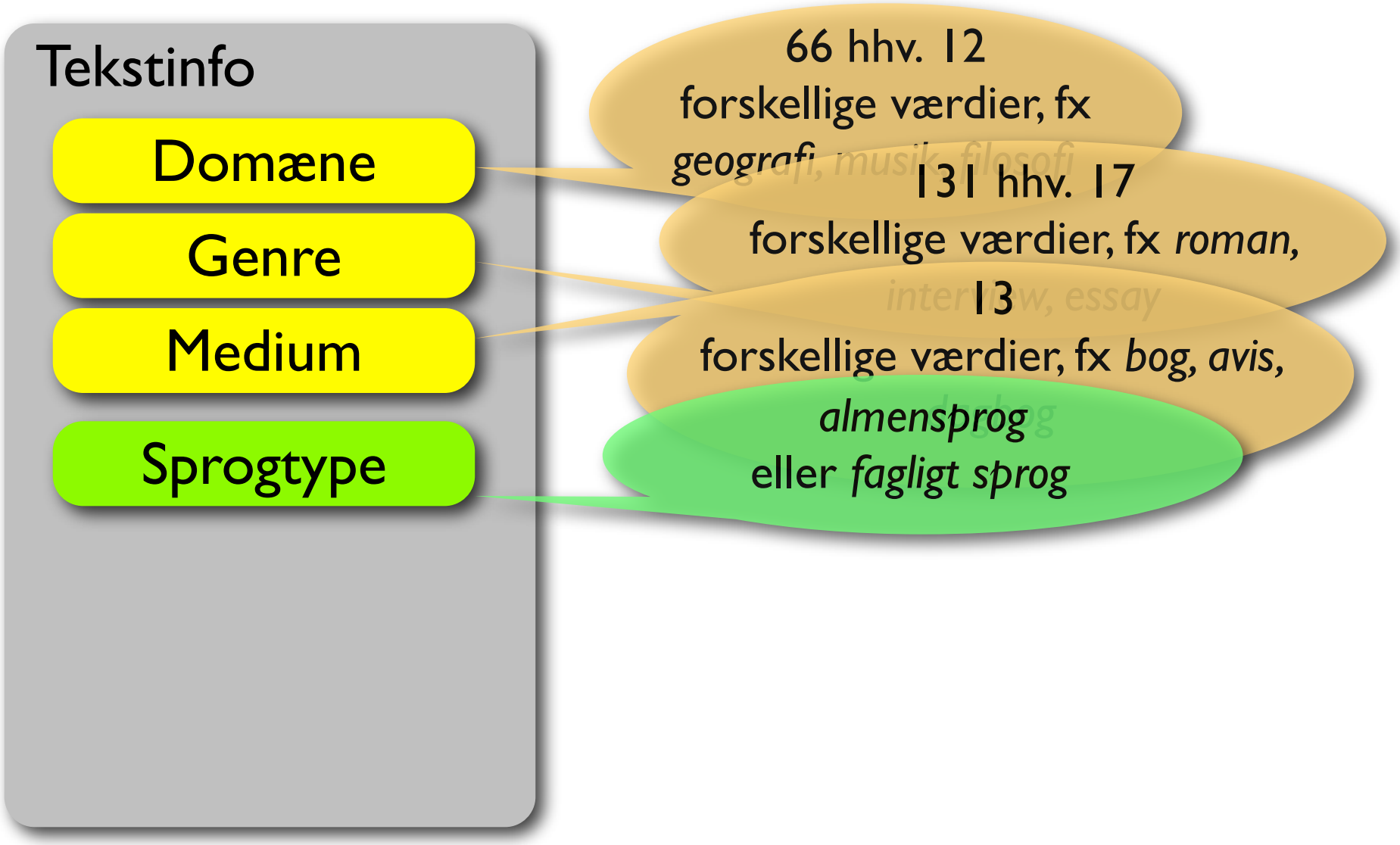
Korpussammensætning



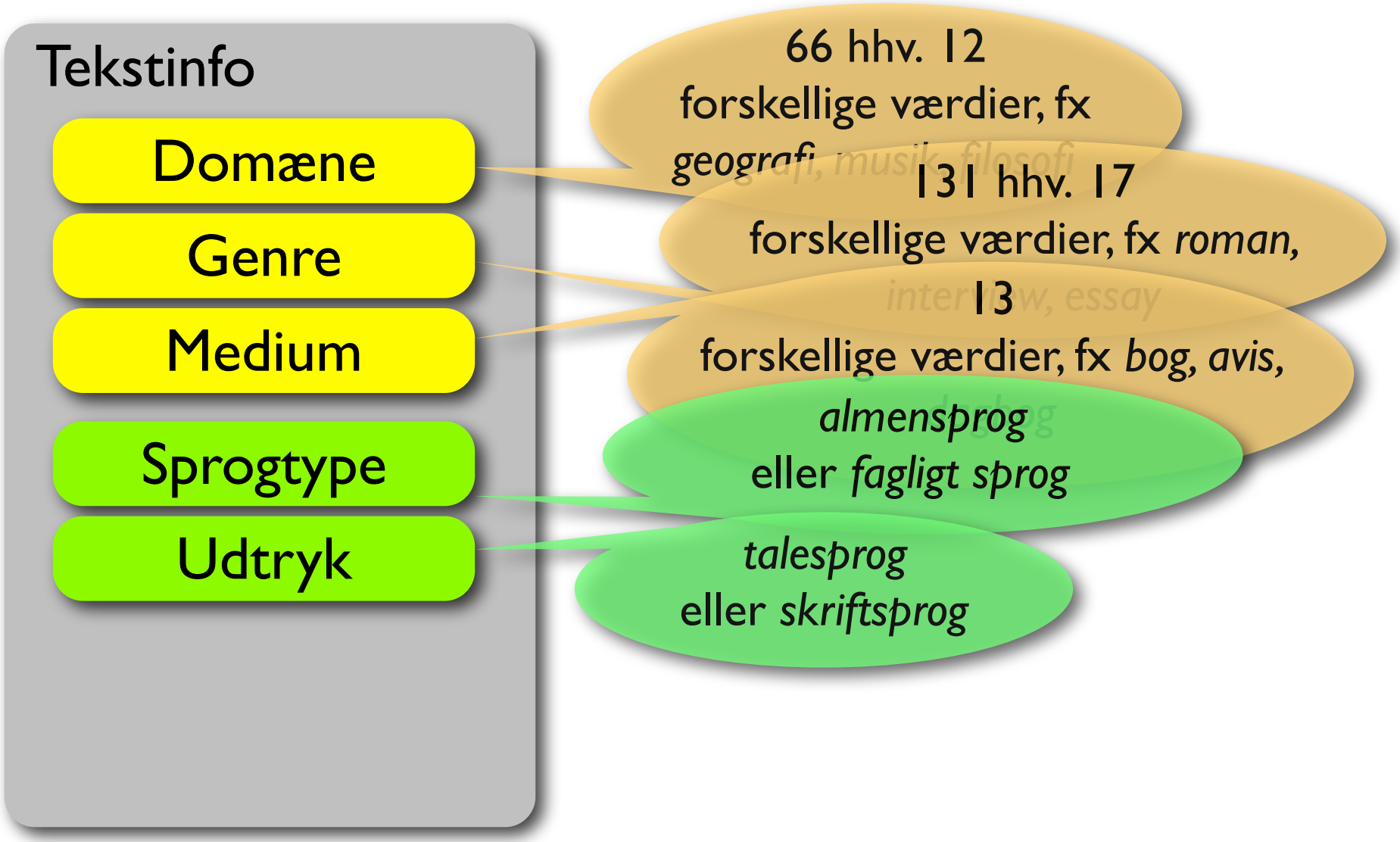
Korpussammensætning



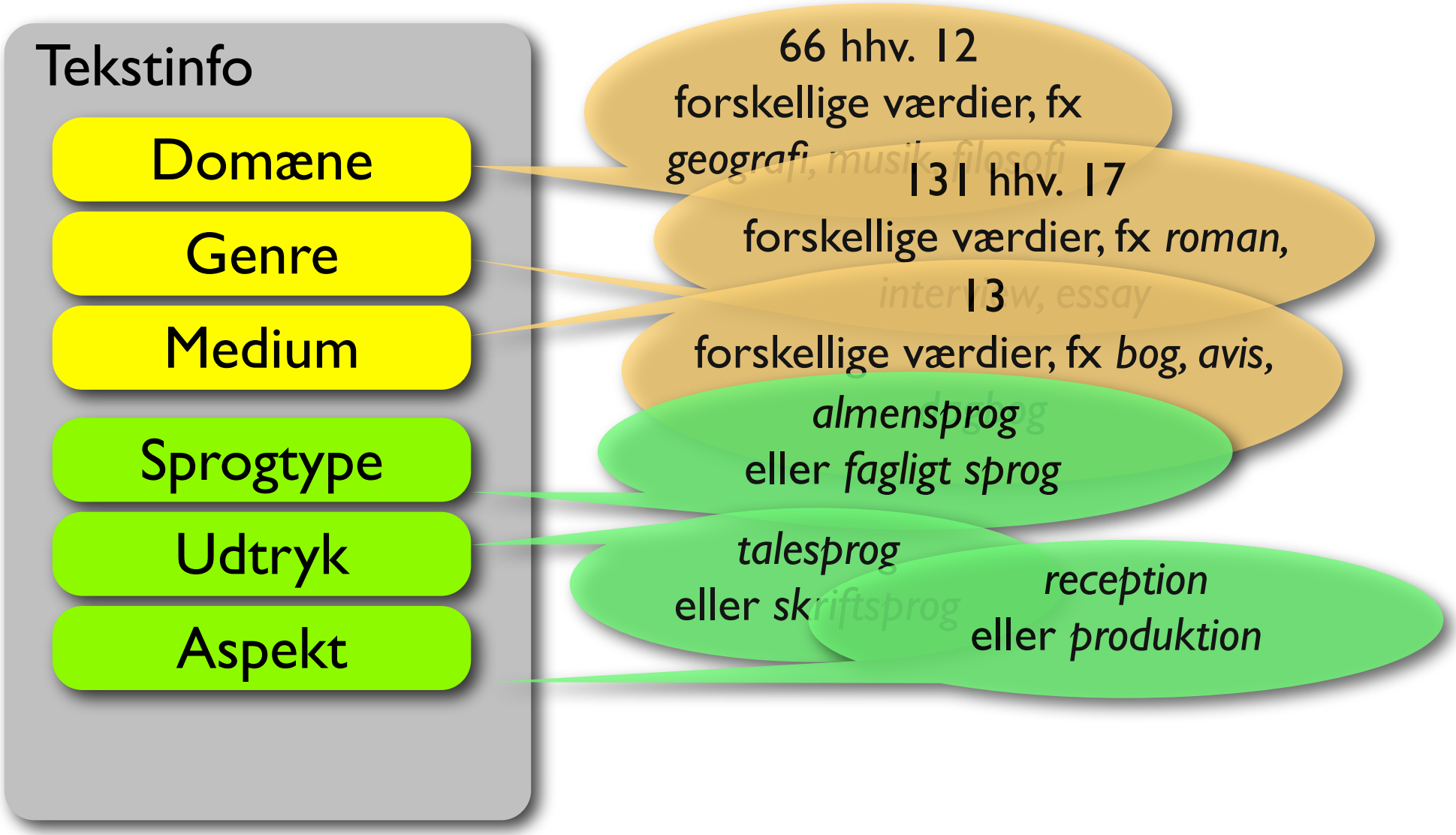
Korpussammensætning



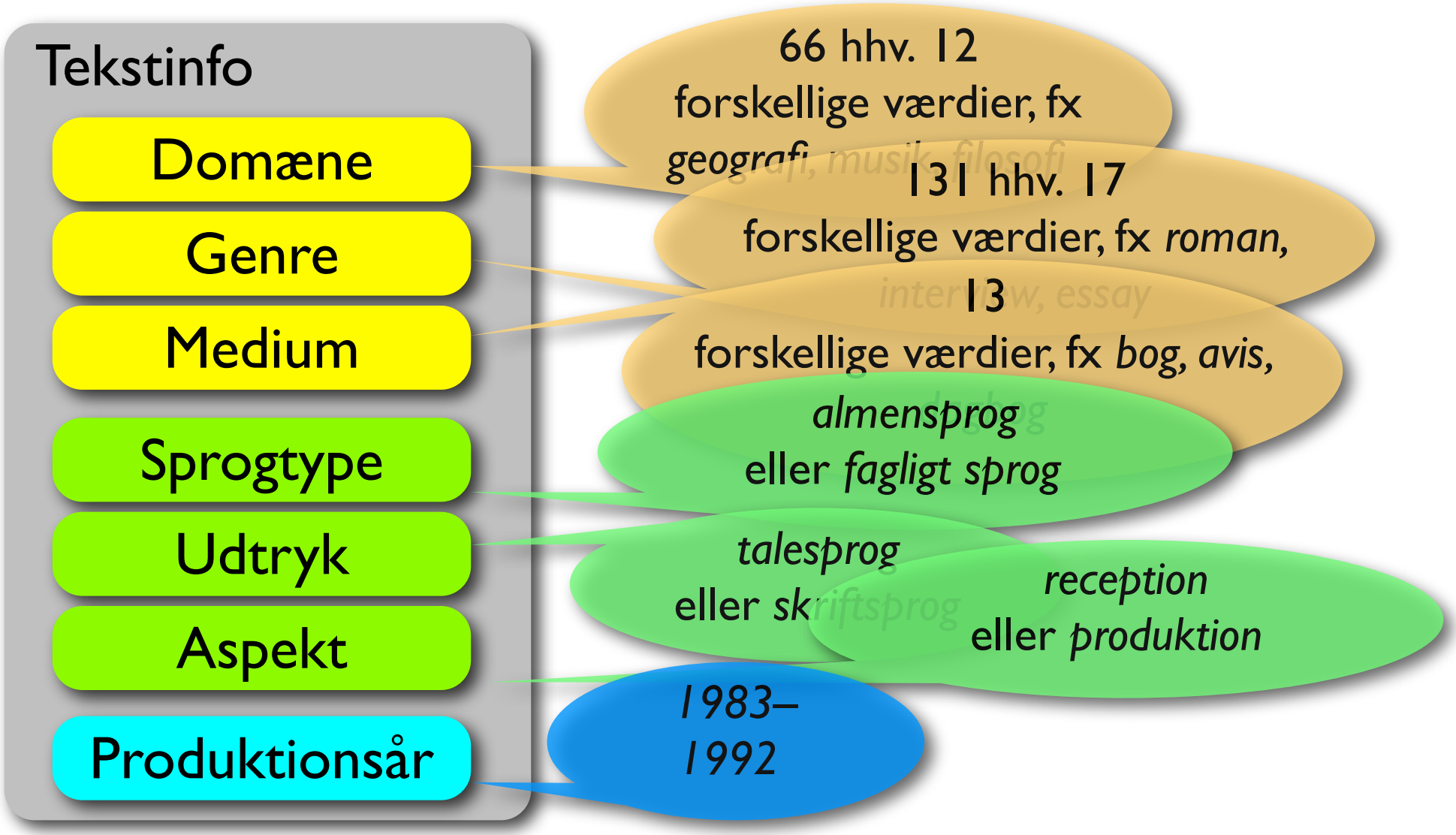
Korpussammensætning



Korpussammensætning



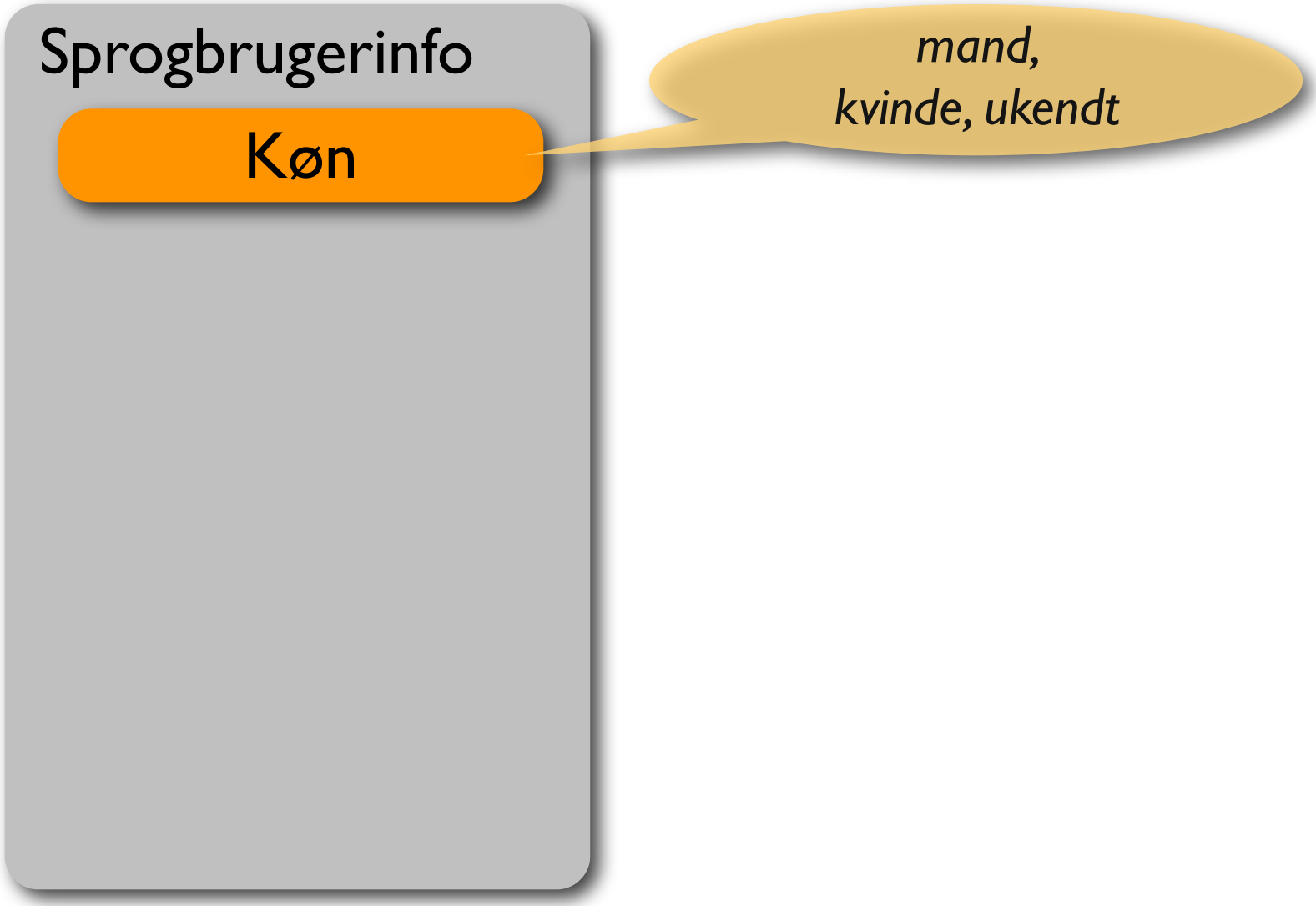
Korpussammensætning



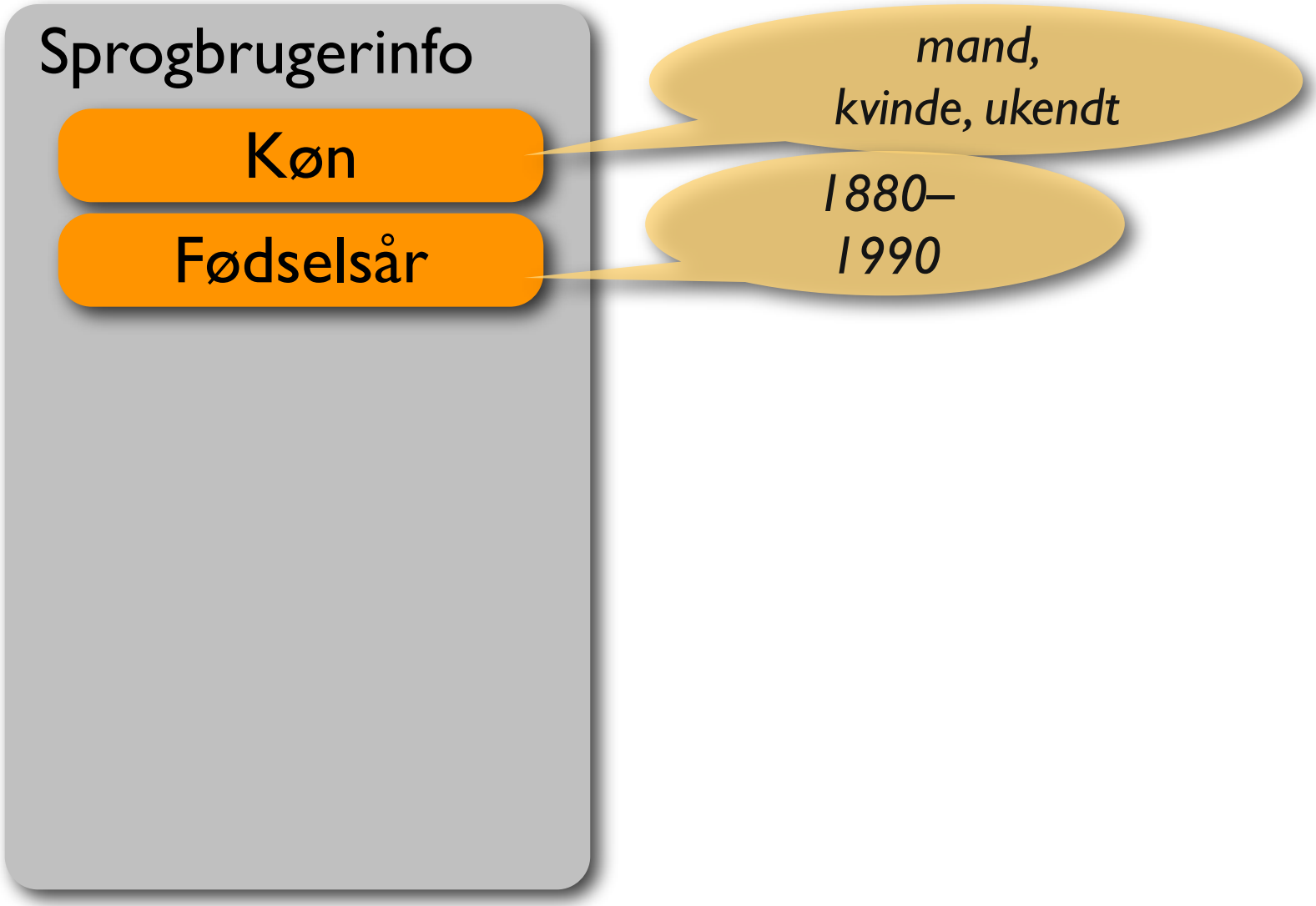
Korpussammensætning

Sprogbrugerinfo

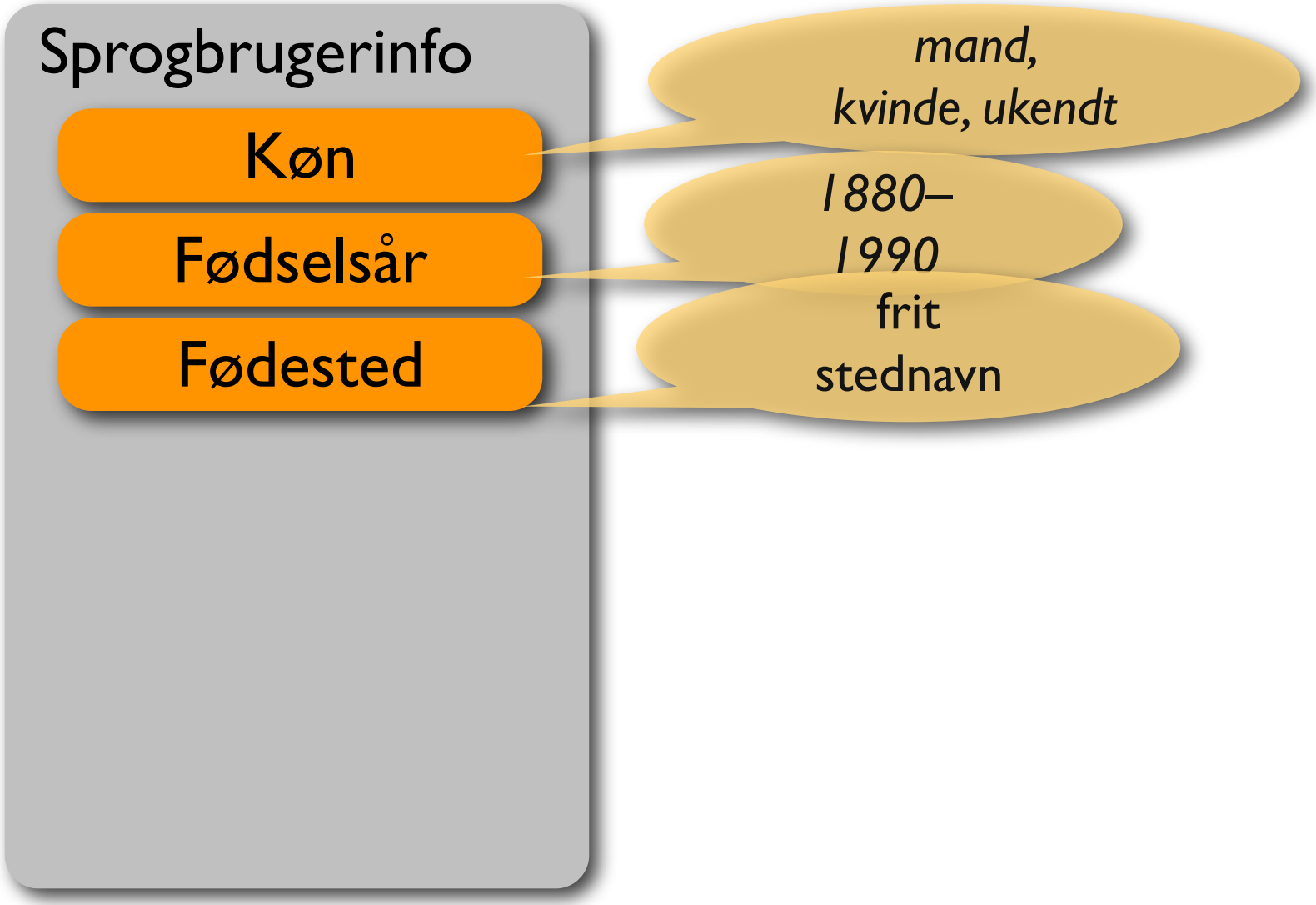
Korpussammensætning



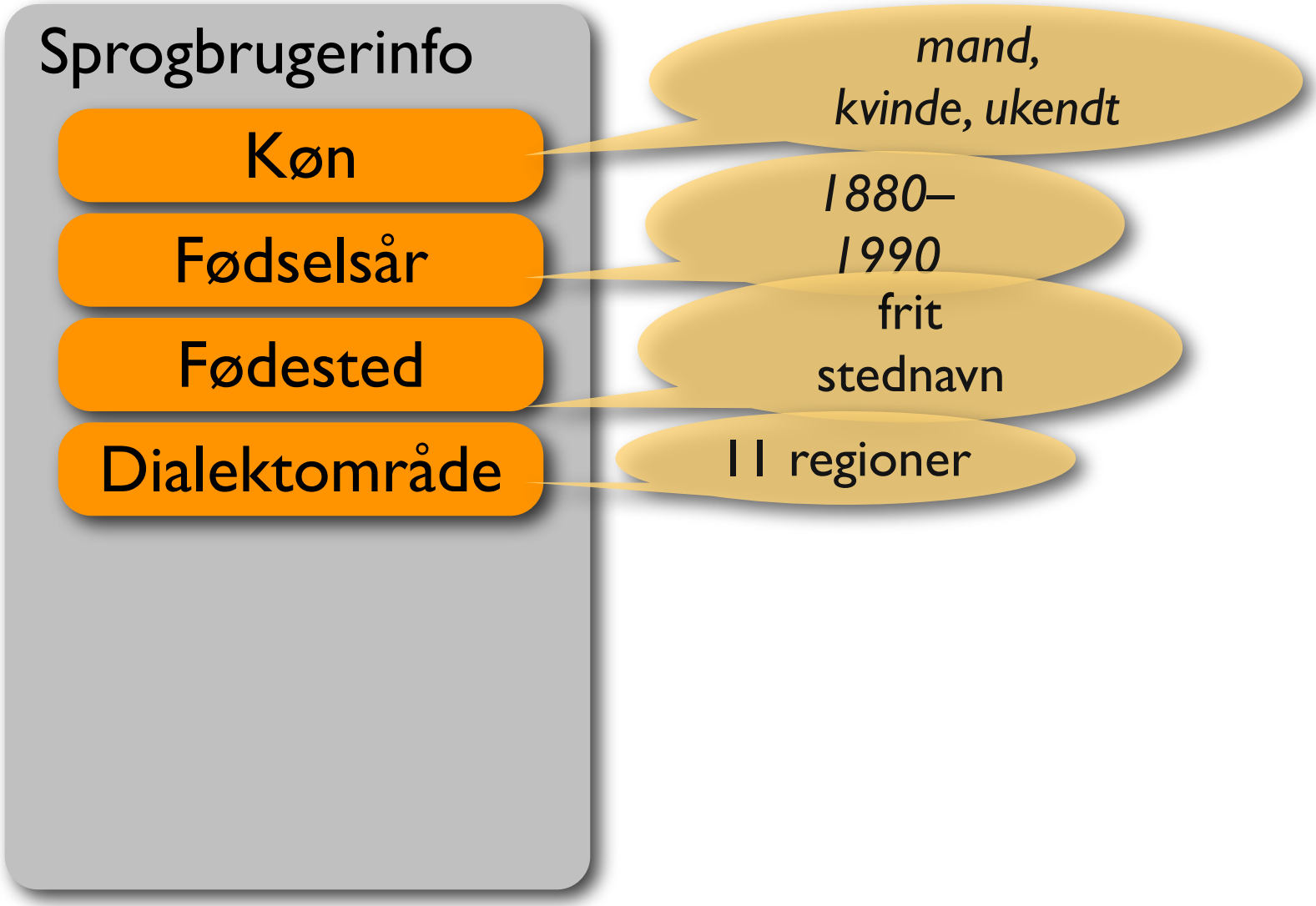
Korpussammensætning



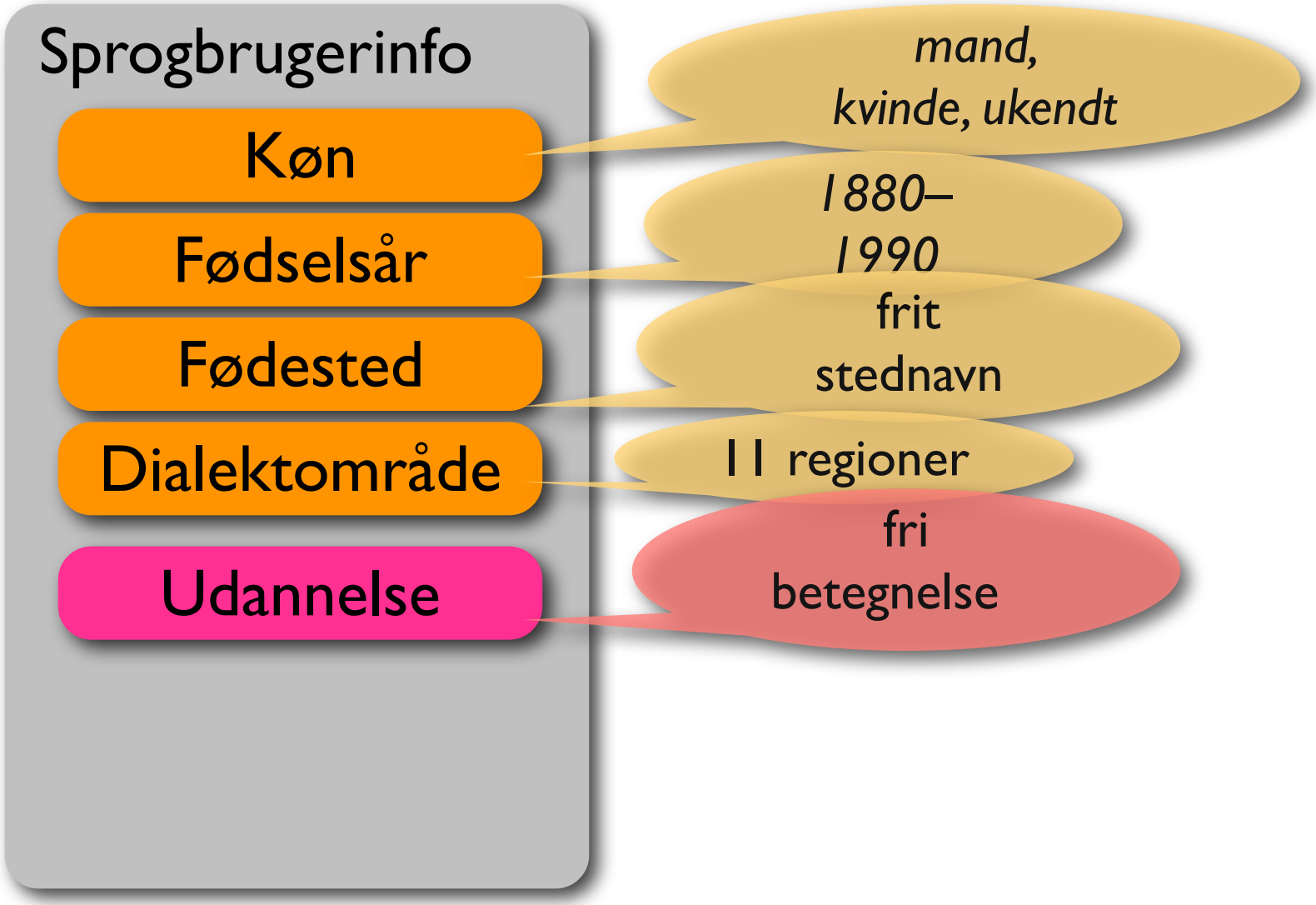
Korpussammensætning



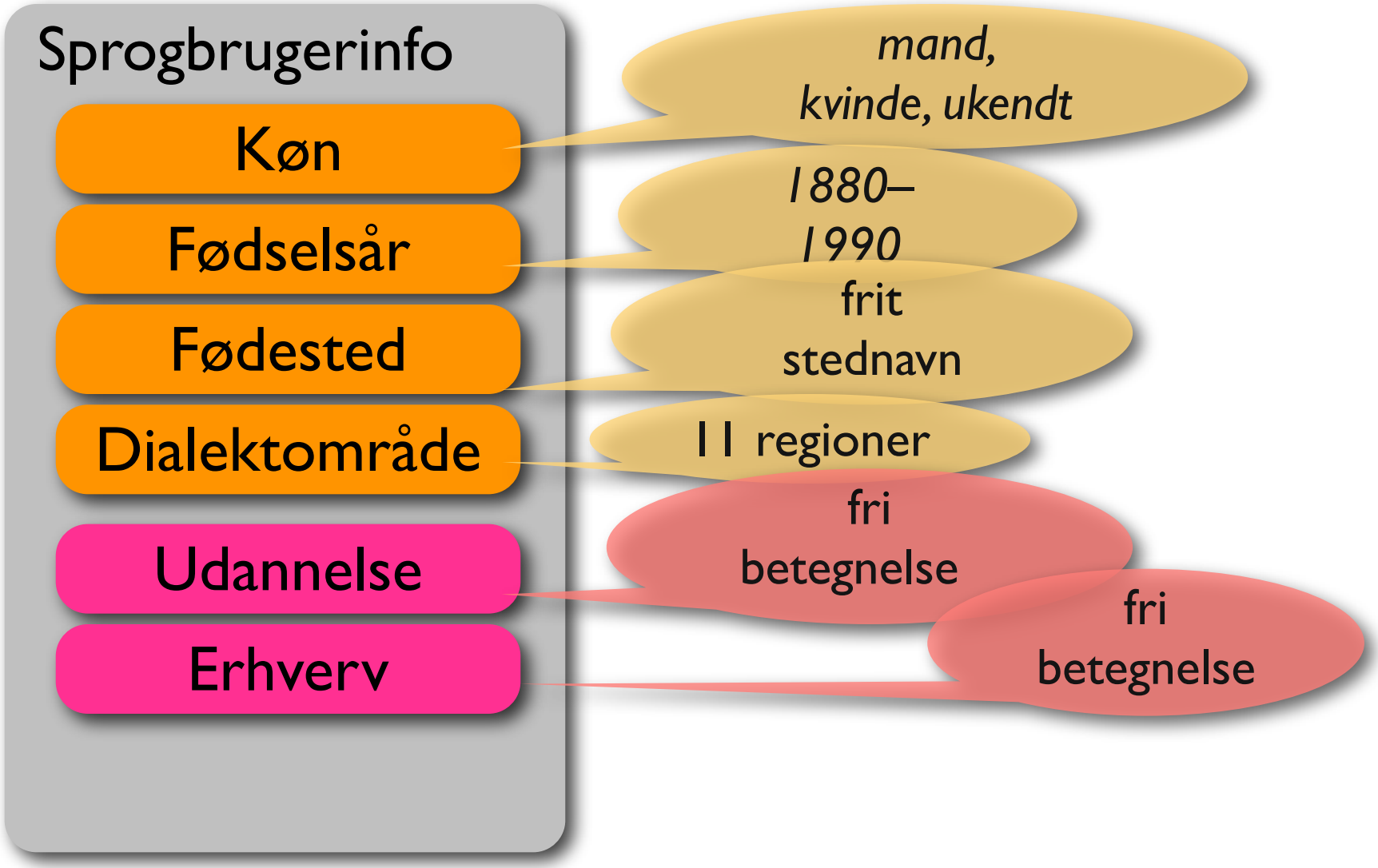
Korpussammensætning



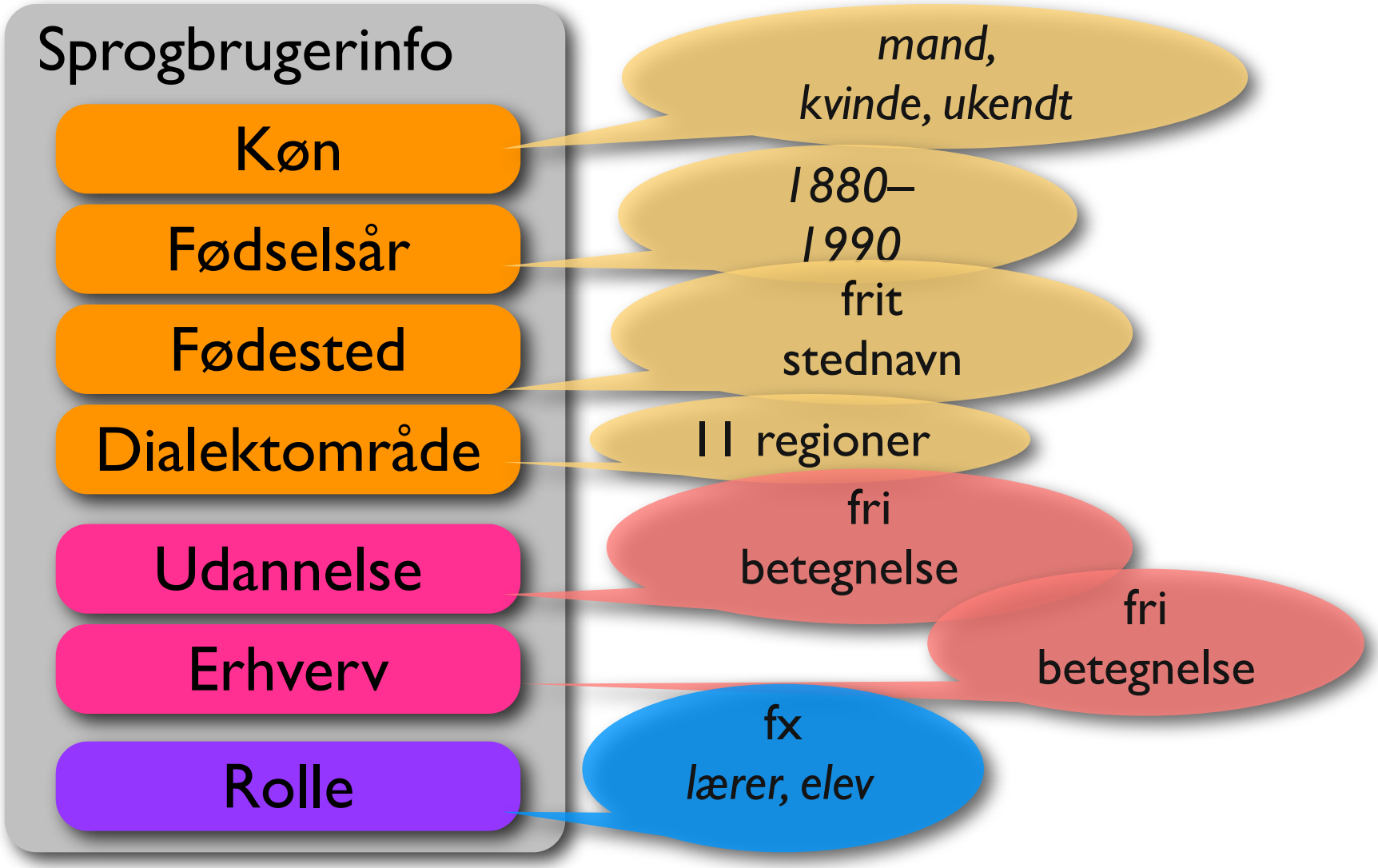
Korpussammensætning



Korpussammensætning

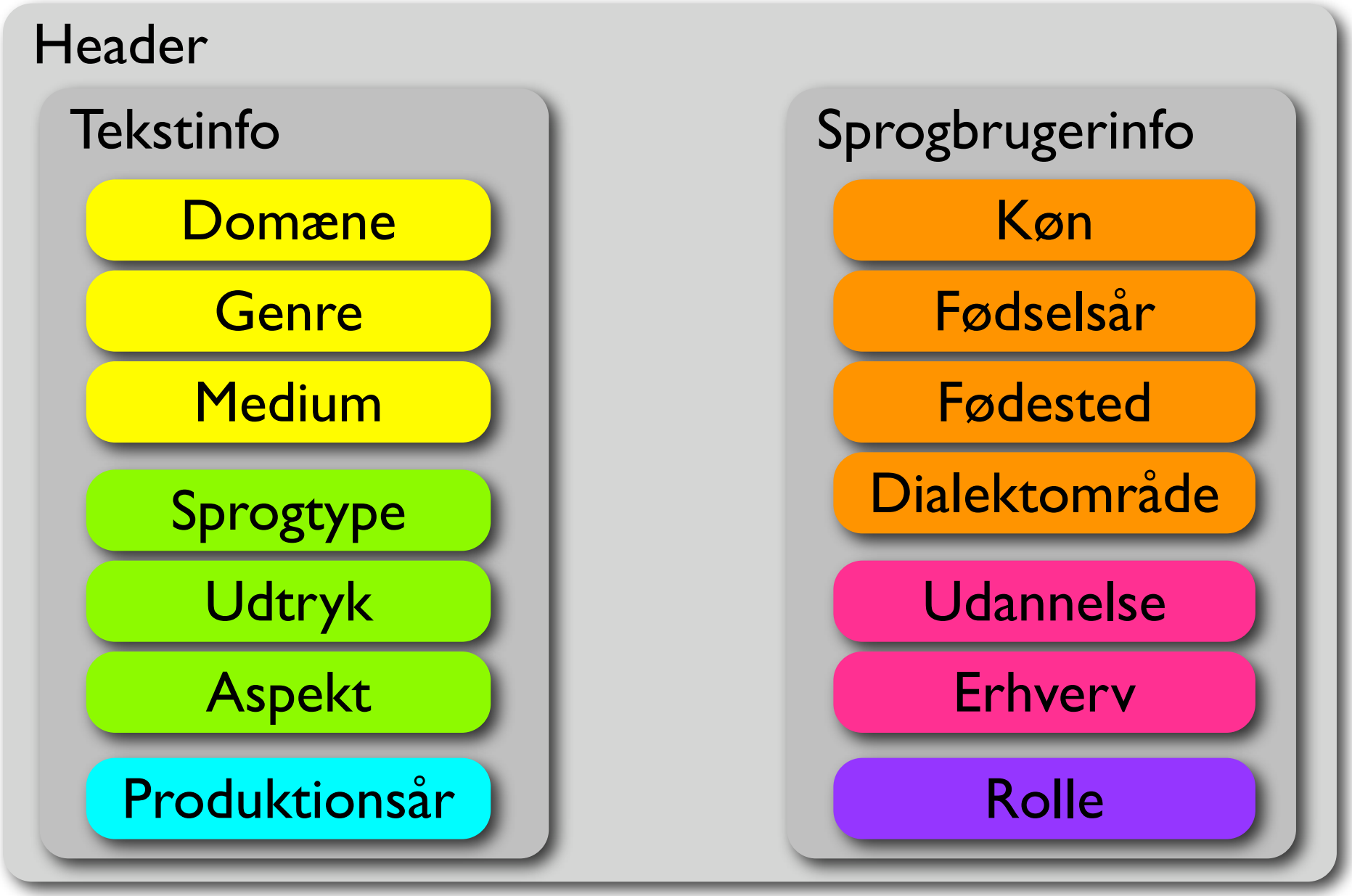


Korpussammensætning



Korpussammensætning

Korpussammensætning



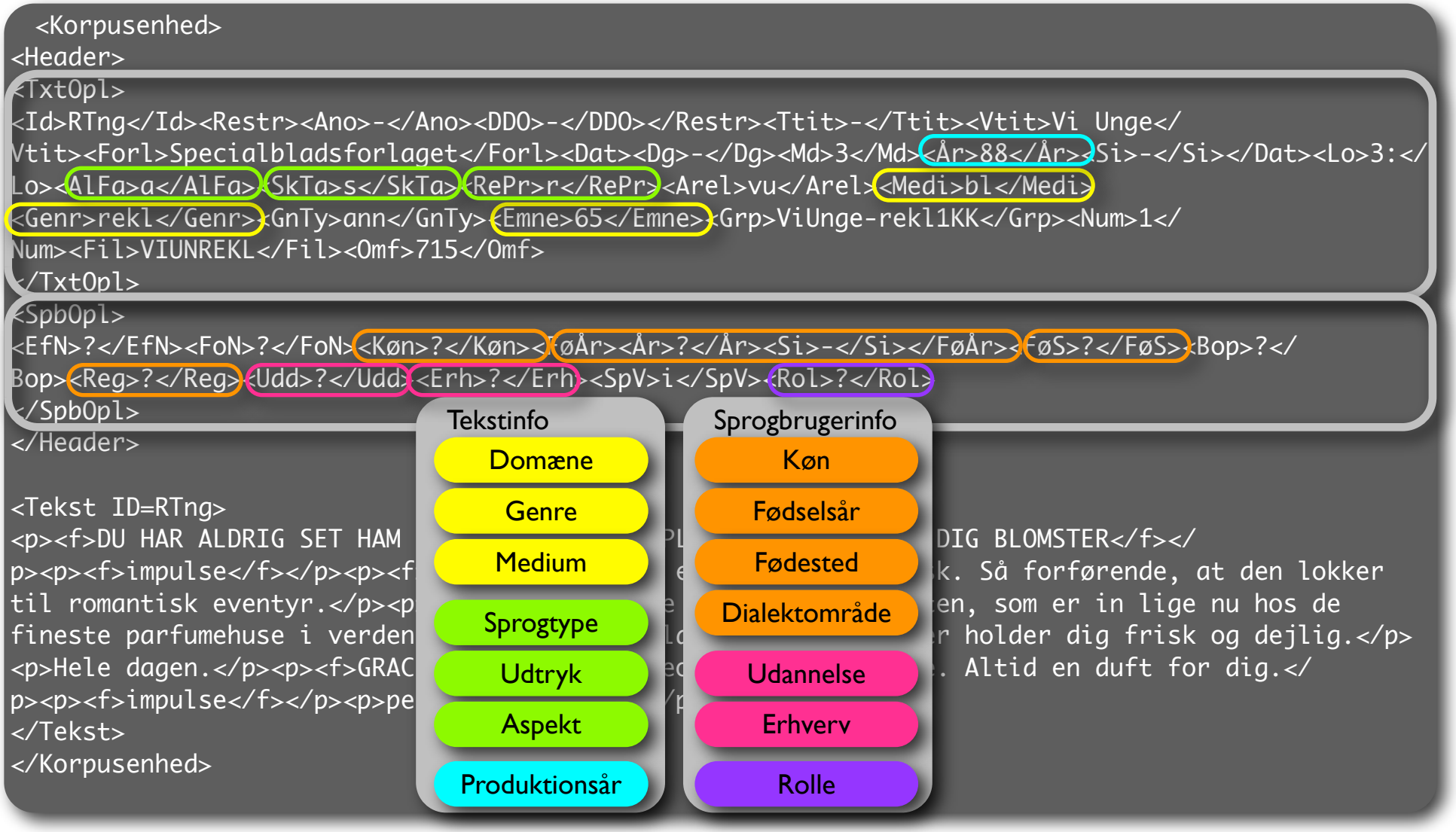
Korpussammensætning

Korpussammensætning

Korpussammensætning

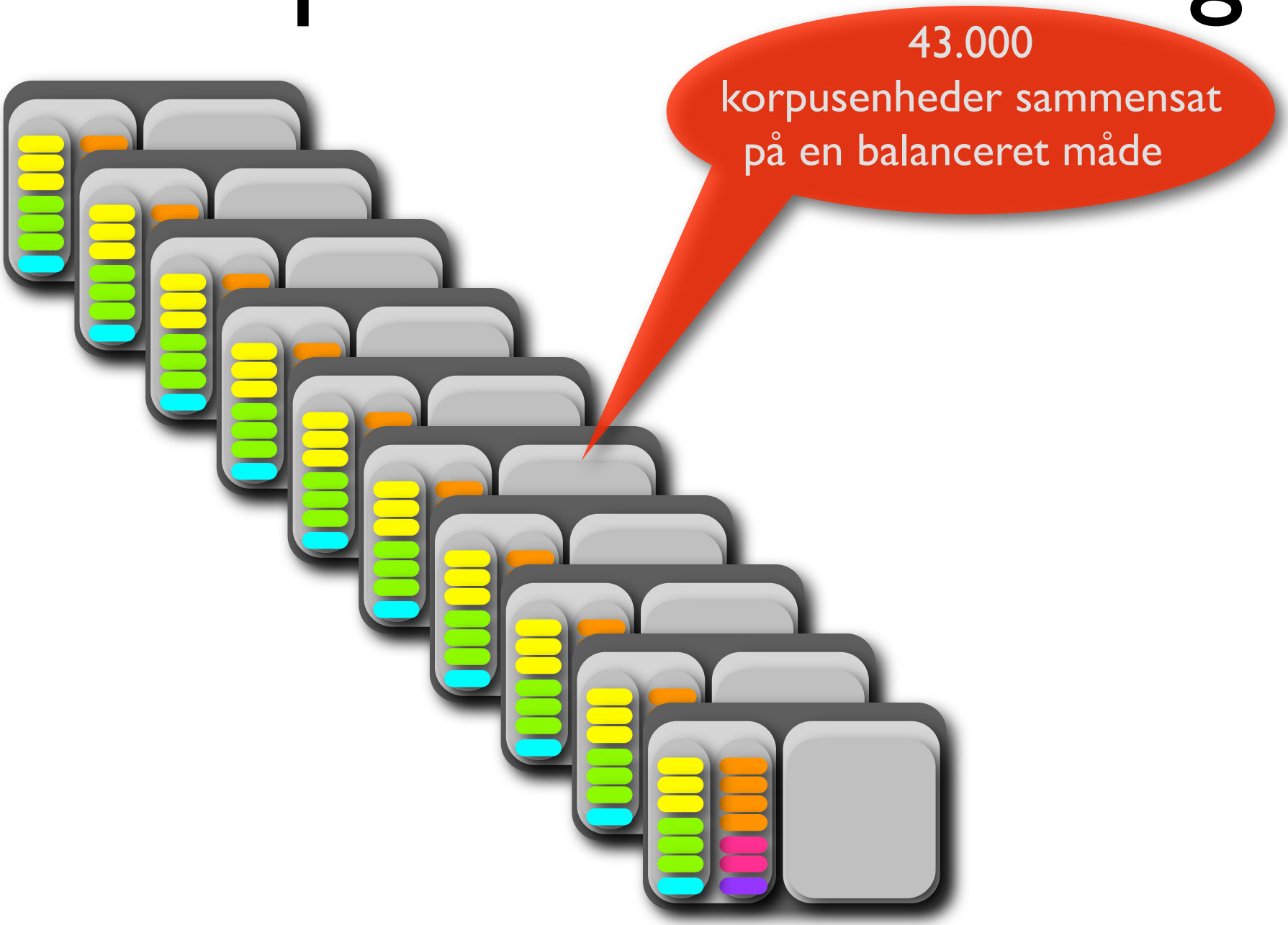
```
 <Korpusenhed>
<Header>
<TxtOpl>
<Id>RTng</Id><Restr><Ano>-</Ano><DDO>-</DDO></Restr><Ttit>-</Ttit><Vtit>Vi Unge</
Vtit><Forl>Specialbladsforlaget</Forl><Dat><Dg>-</Dg><Md>3</Md><År>88</År><Si>-</Si></Dat><Lo>3:</
Lo><AlFa>a</AlFa><SkTa>s</SkTa><RePr>r</RePr><Arel>vu</Arel><Medi>bl</Medi>
<Genr>rekl</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 FØR</f></p><p><f>PLUDSELIG 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



Korpussammensætning

Korpussammensætning



Korpussammensætning

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

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

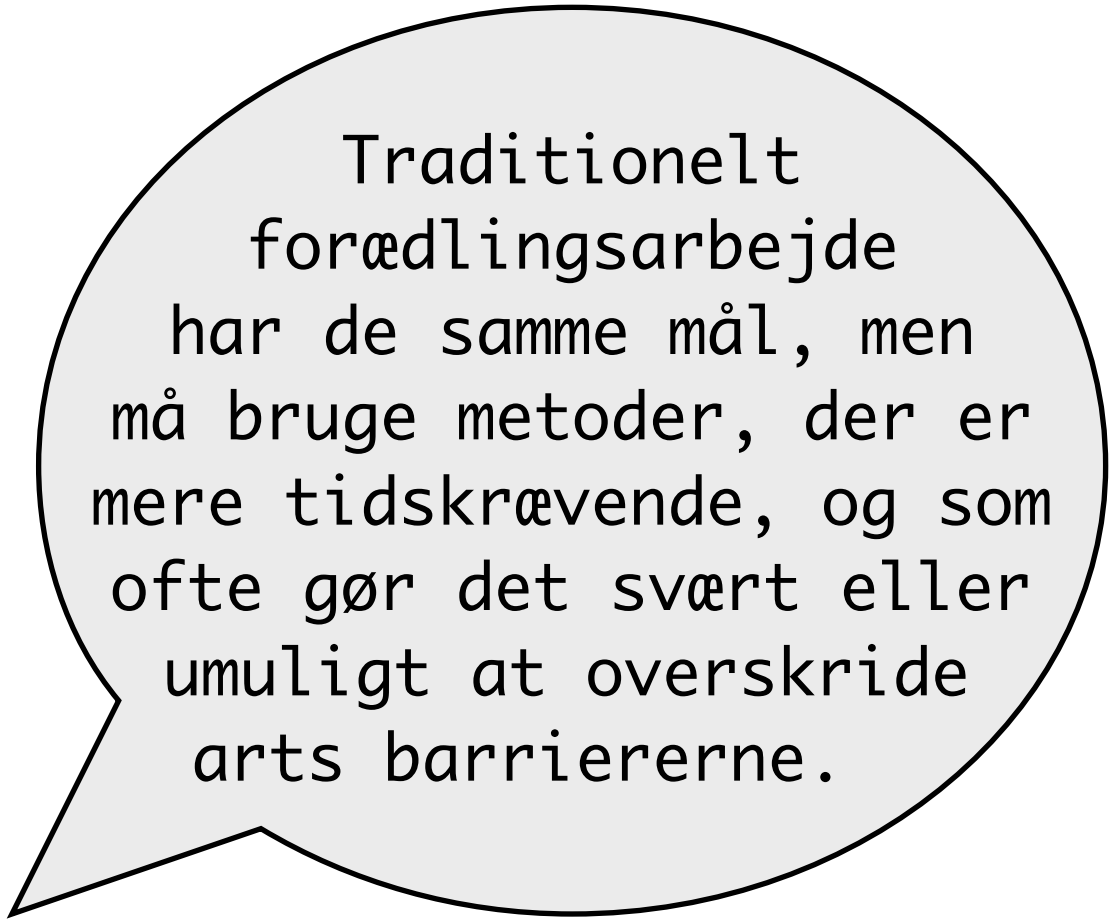
- Udgangspunkt: løbende tekst
  - Token- og sætningsopdeling
  - Lemmatisering
  - Ordklassetagging
  - Syntaktisk parsning
  - Semantisk opmærkning

Abstraktionsniveauer

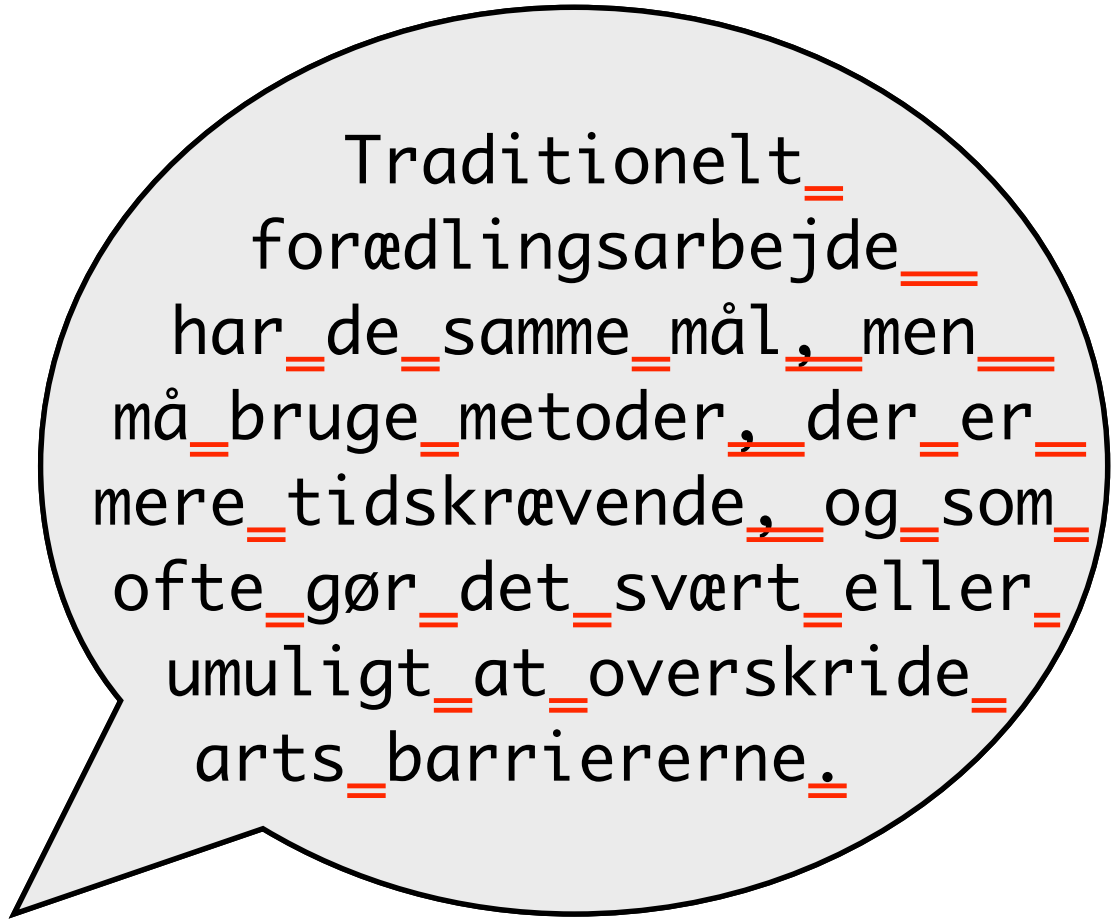
- Udgangspunkt: løbende tekst
  - Token- og sætningsopdeling
  - Lemmatisering
  - Ordklassetagging
  - Syntaktisk parsning
  - Semantisk opmærkning

Disse 3 niveauer
ser vi nærmere på

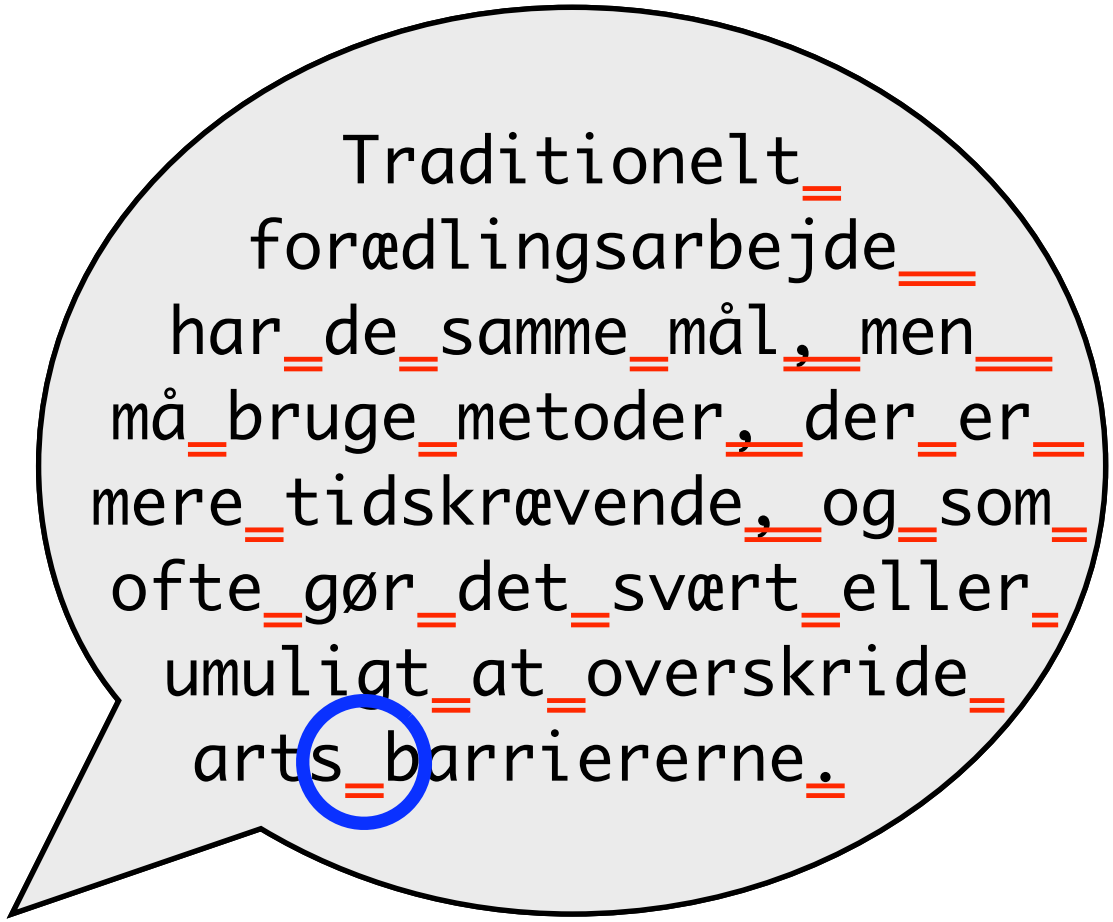
Tokenopdeling



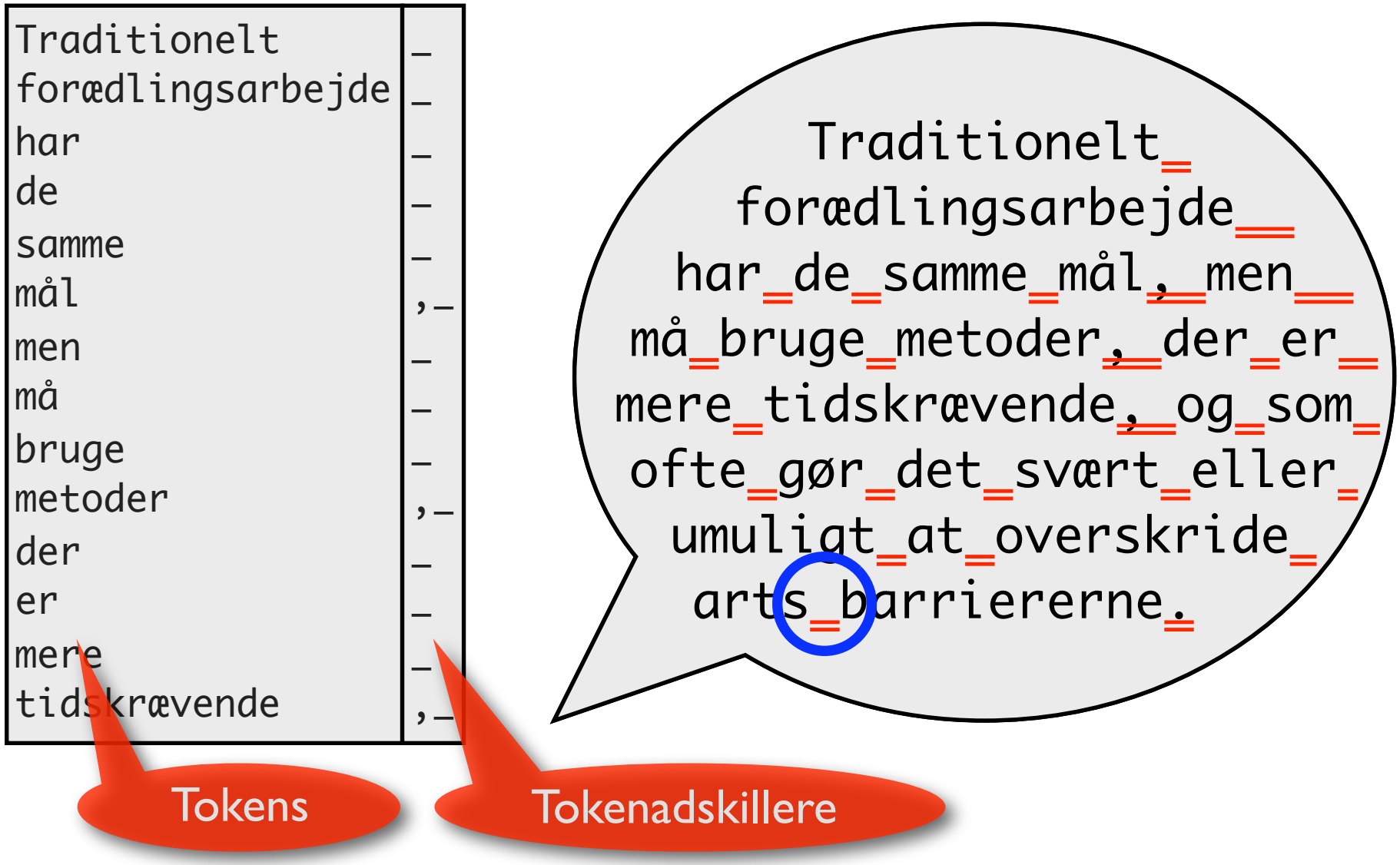
Tokenopdeling



Tokenopdeling



Tokenopdeling



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	,_	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	,_	tidkrævende

Lemmaformer (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	_	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

Ordklassetagging forudsætter et fuldformsleksikon og en disambigueringsrutine

Ordklassetags

Bøjningstags

Tekstformat

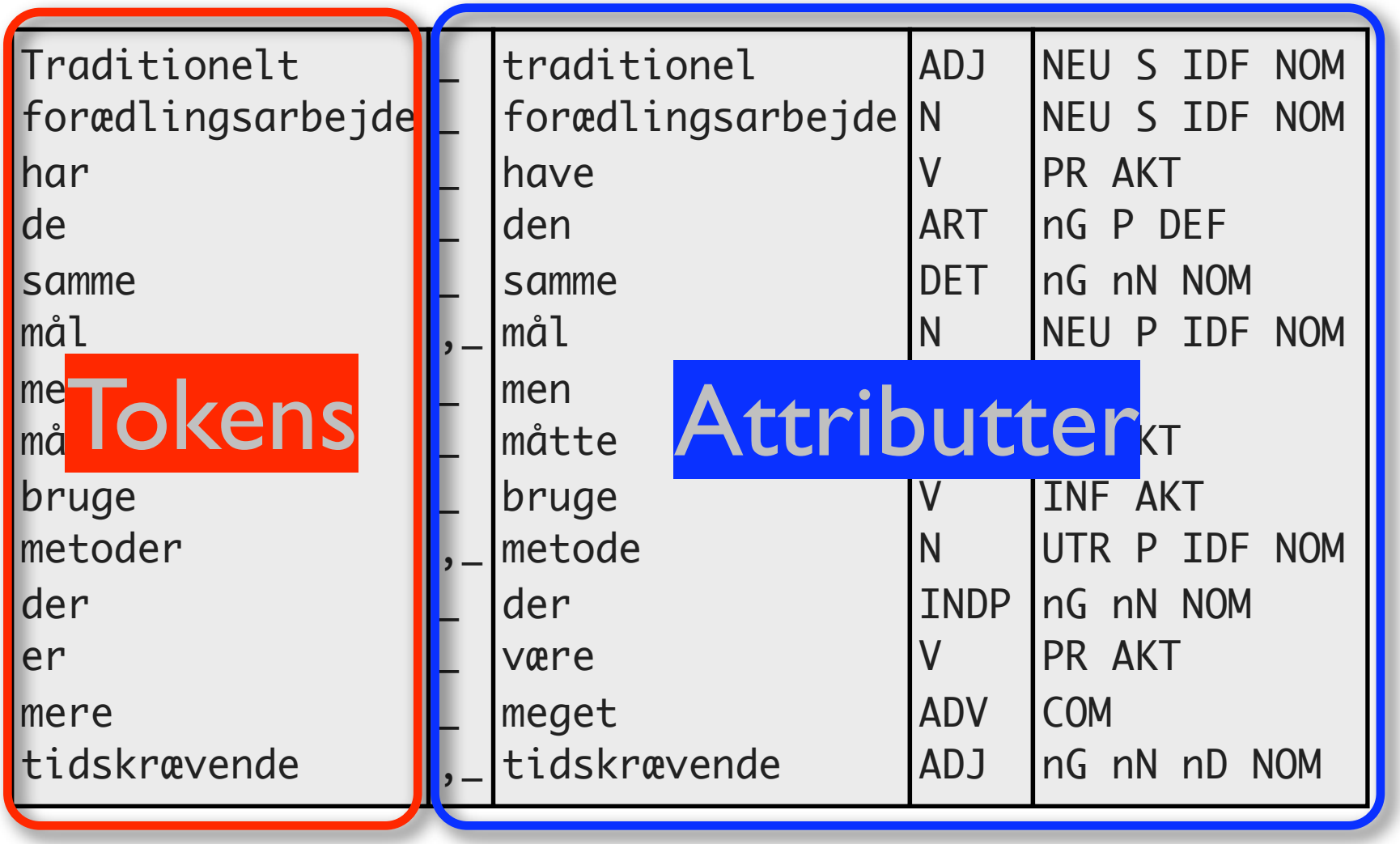
Traditionelt	_	traditionel	ADJ	NEU S IDF NOM
forædlingsarbejde	_	forædlingsarbejde	N	NEU S IDF NOM
har	_	have	V	PR AKT
de	_	den	ART	nG P DEF
samme	_	samme	DET	nG nN NOM
mål	,_	mål	N	NEU P IDF NOM
men	_	men	KC	
må	_	måtte	V	PR AKT
bruge	_	bruge	V	INF AKT
metoder	,_	metode	N	UTR P IDF NOM
der	_	der	INDP	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	_	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		
må	_	måtte		KT
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

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?

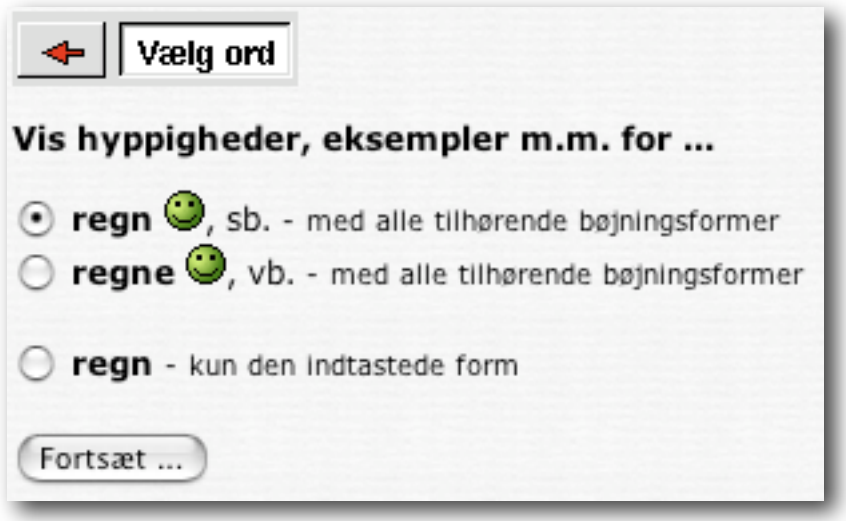
Søgning på lemma

Søgning på lemma

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

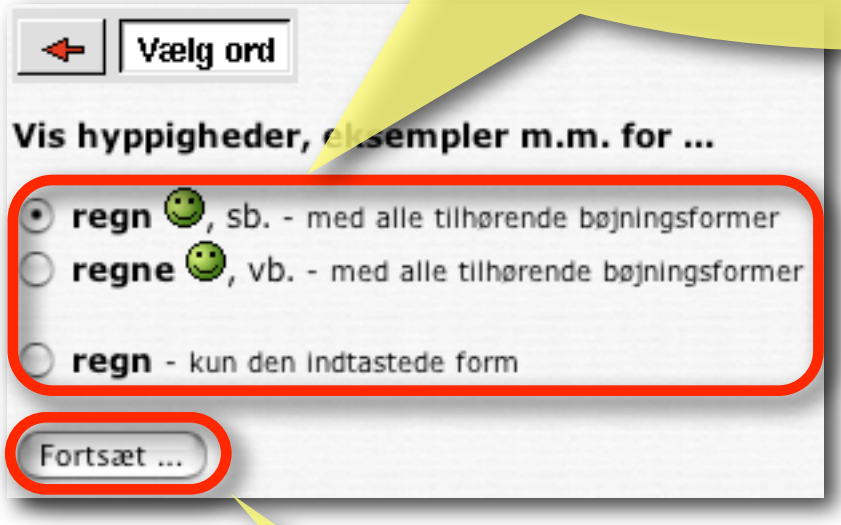
Klik
her

Søgning på lemma



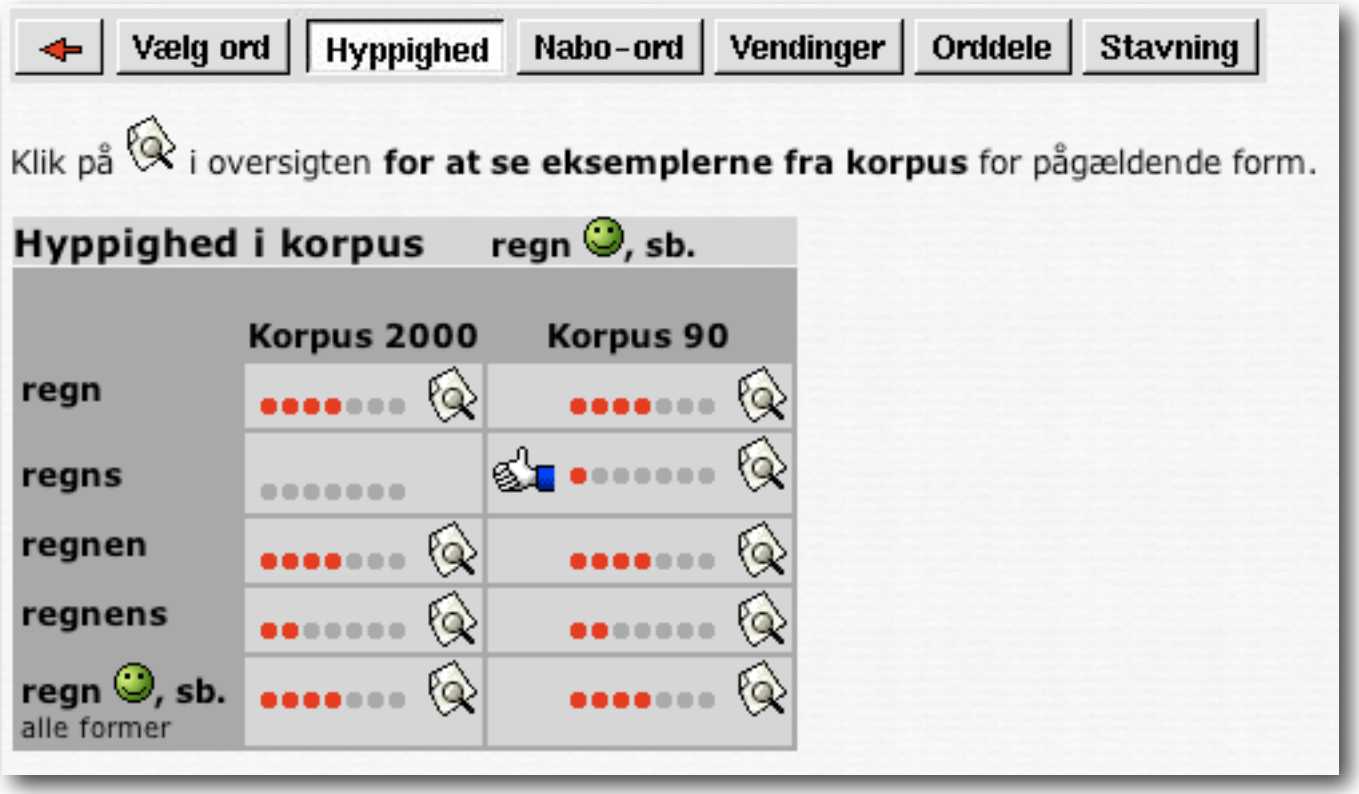
Søgning på lemma

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

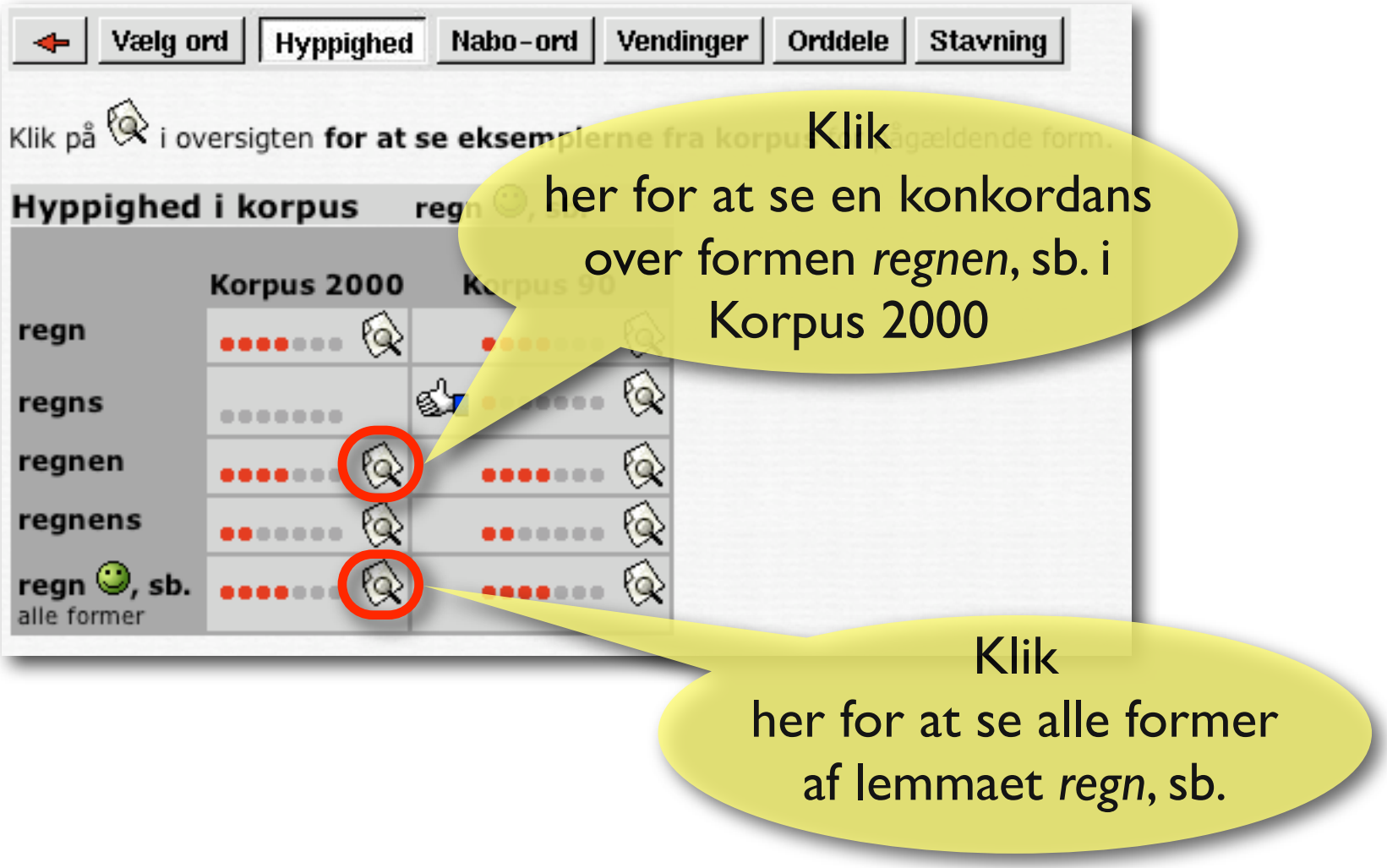


Klik
her

Søgning på lemma



Søgning på lemma



Søgning på lemma

Søgning på lemma

Resultatet er en KWIC-konkordans, sorteret

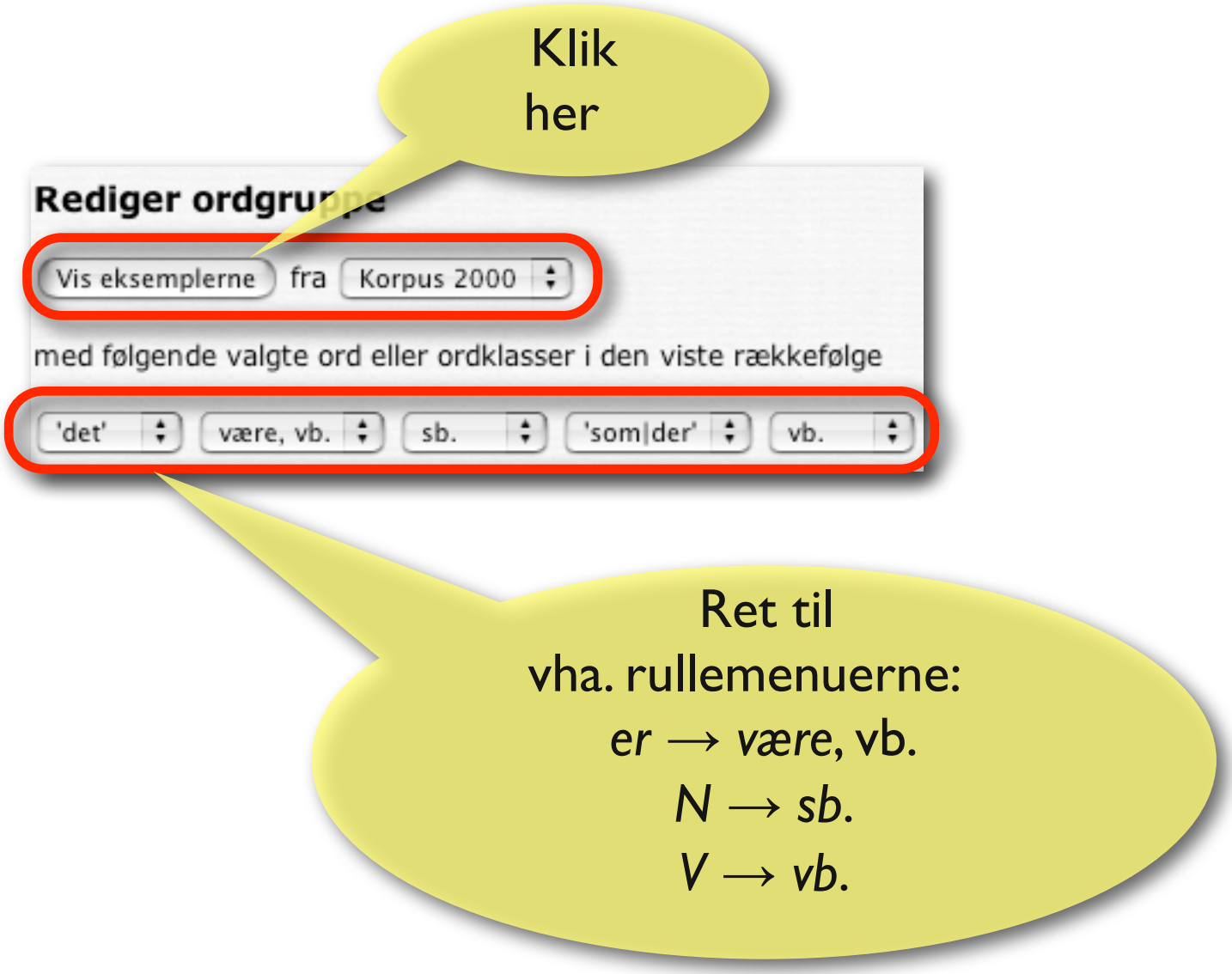
KWIC
= keyword in context

Sætningskløvning

Sætningskløvning

Sætningskløvning

Sætningskløvning



Sætningskløvning

Sætningskløvning

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*?

Hvad er *ordnet.dk*?

Hvad er *ordnet.dk*?

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