

Emne

“Korpus 2010” DK-CLARIN

Fortæller

Jørg Asmussen
Det Danske Sprog- og Litteraturselskab

Tid & Sted

Torsdag, 24. maj 2012
NoSSinK, Dansk Sprognævn

Mere

Denne præsentation findes under
<http://korpus.dsl.dk/staff/ja/pres/nossink.pdf>

WP2.1 Reference corpus of general language

Some corpora for Danish do exist, but there is a need for intensifying the compilation of corpus material both in terms of quantity and with respect to continuity in order to secure the diachronic aspect. The project will collect at least 15 million words of Danish text per year.

Material will mainly be taken from newspapers and periodicals. Some basic methods and tools for automatically collecting, structuring and annotating text will be developed, using KB's records of the entire .dk domain, harvested 4 times per year. Material will be collected taking the copyright into consideration and all the collected text will be made available to the research community without restrictions in so far as copyrights permit, cf. WP1.

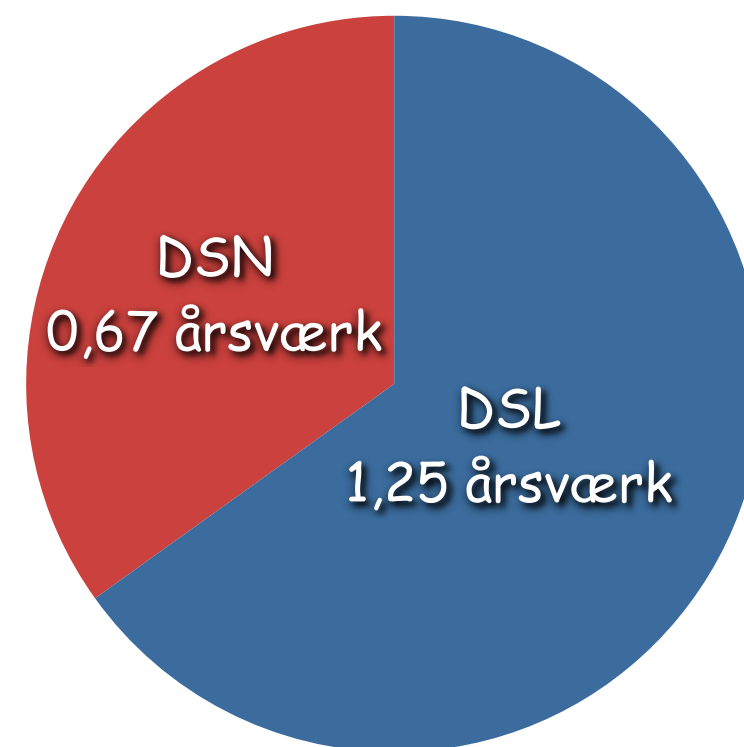
Flere oplysninger:
<http://clarin.dsl.dk>

* Deltagere:

- * Jørg Asmussen, DSL
- * Jakob Halskov, DSN
- * Liisa Theilgaard, DSL

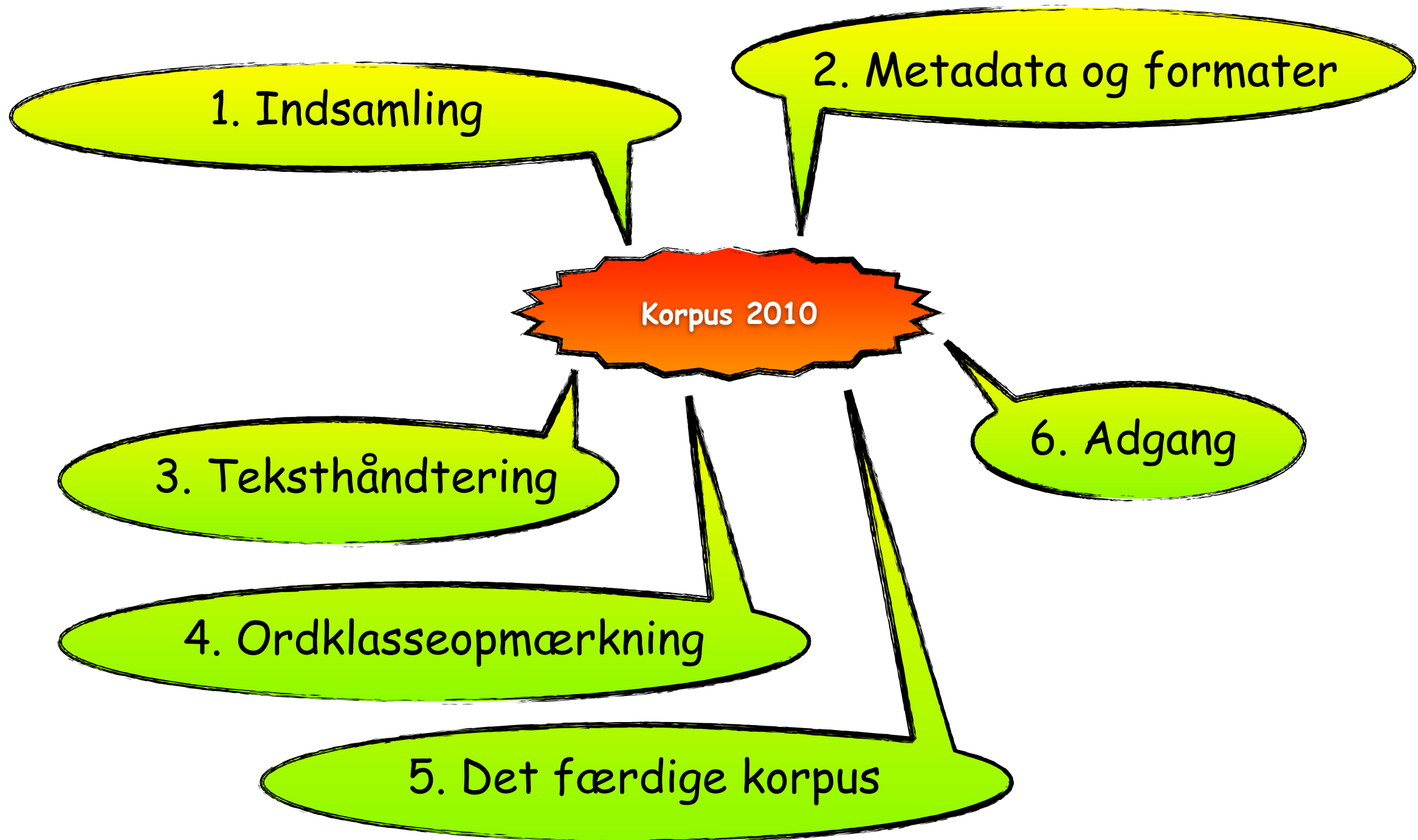
* 1 million kroner:

- * DSL 700.000
- * DSN 300.000



Arbejdsopgaver ...

Arbejdsopgaver = Program



Indsamling ...

1. Indsamling

- * Taping fra KB's arkiv over dk-domænet
- * Løbende Infomedia-indsamling
 - * DSL siden 2005
 - * DSN siden 2008
- * "Frit" materiale (DSN):
 - * Folketinget
 - * Wikipedia
- * CLARIN-aftale (DSN):
 - * Bo Bedre, Smag & Behag, Se & Hør, Irma, Liberalt Overblik, Socialdemokraten, UNG, Idénnyt, Vi Unge, Tjeck, Samvirke, Ud & Se
- * Fælles-aftale:
 - * Via Infomedia (DSL): Helse, Magasinet Ejendom, Juristen, djøfbladet, Hus Forbi, Ældre Sagen, Højskolebladet, Folkeskolen, Samfundsøkonomen, penge.dk
 - * Via webbet (DSN): blogsbjerg.com, skauogco.blogspot.com, jarlcardua.dk, Lifli, Bentes Blog, blogbogstaver.dk, selvhenter.org, Mangamania, clionline.dk

dk-domænet...

1. Indsamling: tappe dk-domænet?



netarkivet.dk indsamler og bevarer den danske del af internettet

Adgang til Netarkivet

Netarkivet kan kun benyttes til videnskabelige formål. Det skyldes, at Netarkivet også indeholder følsomme personoplysninger. Følgende kan få adgang:

- Forskere og ph.d.-studerende kan få online adgang til arkivet.
- Andre vil på et senere tidspunkt kunne benytte arkivet på Det Kongelige Bibliotek og Statsbiblioteket, såfremt formålet med benyttelsen er videnskabelige studier.
- Herudover kan du få adgang til dit eget websted, for så vidt det ikke indeholder oplysninger om andre personer.

Sådan får du adgang

Adgang til Netarkivet gives kun efter ansøgning. [Se retningslinjer her](#).

Du skal udfylde denne [ansøgererklæring](#) og sende den med post til:

Eva Fønss-Jørgensen
Victor Albecks Vej 1
Statsbiblioteket
8000 Aarhus C

Lad os se på en fordeling ...

1. Indsamling: CLARIN-aftale

Vedr. tekster til DK-CLARIN-projektet

DSB giver hermed tilladelse til at tekstindholdet af udvalgte numre af bladet *Ud & Se* optages i en samling elektroniske tekster under opbygning i projektet DK-CLARIN, der som mål har at skabe en dansk it-infrastruktur som skal rumme skrevne og talte tekster og visuelle data i et sammenhængende og systematisk digitalt lager som gøres tilgængeligt for forskning.

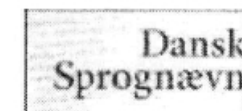
Tekstmaterialet stilles til rådighed for DK-CLARIN og vil udelukkende blive tilgængelige for forskere ved universiteter og andre forskningsinstitutioner samt udvalgte studerende som deltager i relevant undervisning. Tekstmaterialet vil udelukkende blive anvendt til ikke-kommercielle formål, og adgang til tekstmaterialet vil blive begrænset til forskere og studerende ved hjælp af et password-beskyttet internet-login.

Desuden gives der tilladelse til at teksterne kan vises for offentligheden, på følgende vilkår: Brugere får ikke adgang til teksterne i deres helhed, men kun til begrænsede tekstuddrag (op til ca. 100 ord på hver side af det søgte ord), der kun kan give oplysning om sprogbrugen.

Denne tilladelse gives under forudsætning af at ovenstående principper overholdes.

Med venlig hilsen

9. marts 2010



DSL/DSN-fællesaftale...

1. Indsamling: DSL/DSN-fællesaftale

Mellem

[navn]

[adresse]

[CVR-nr.]

(herefter kaldet Tekstleverandøren)

og

Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
CVR-nr. 59188917

og

Dansk Sprognævn
H.C. Andersens Boulevard 2
1553 København V
CVR-nr. 59190628

(herefter samlet kaldet DSL/DSN)

er der dags dato indgået følgende aftale om at stille tekstmateriale (herefter kaldet Tekstmaterialet) til rådighed til brug for DSL/DSN's digitale tekstsamlinger.

Tekstmaterialet er specificeret i Bilag 1 nedenfor.

Konklusionen blev...

1. Indsamling

- * Taping fra KB's arkiv over dk-domænet

Måtte opgives

- * Løbende Infomedia-indsamling

- * DSL siden 2005

- ~~* DSN siden 2008~~

Overvejende komplementære

- * "Frit" materiale (DSN):

- * Folketinget

- * Wikipedia

Aftale-status uvis

Infomedia

- * CLARIN-aftale (DSN):

- * Bo Bedre, Smag & Behag, Se & Hør, Irma, Liberalt Overblik, Socialdemokraten, UNG, Idényt, Vi Unge, Tjeck, Samvirke, Ud & Se

- * Fælles-aftale:

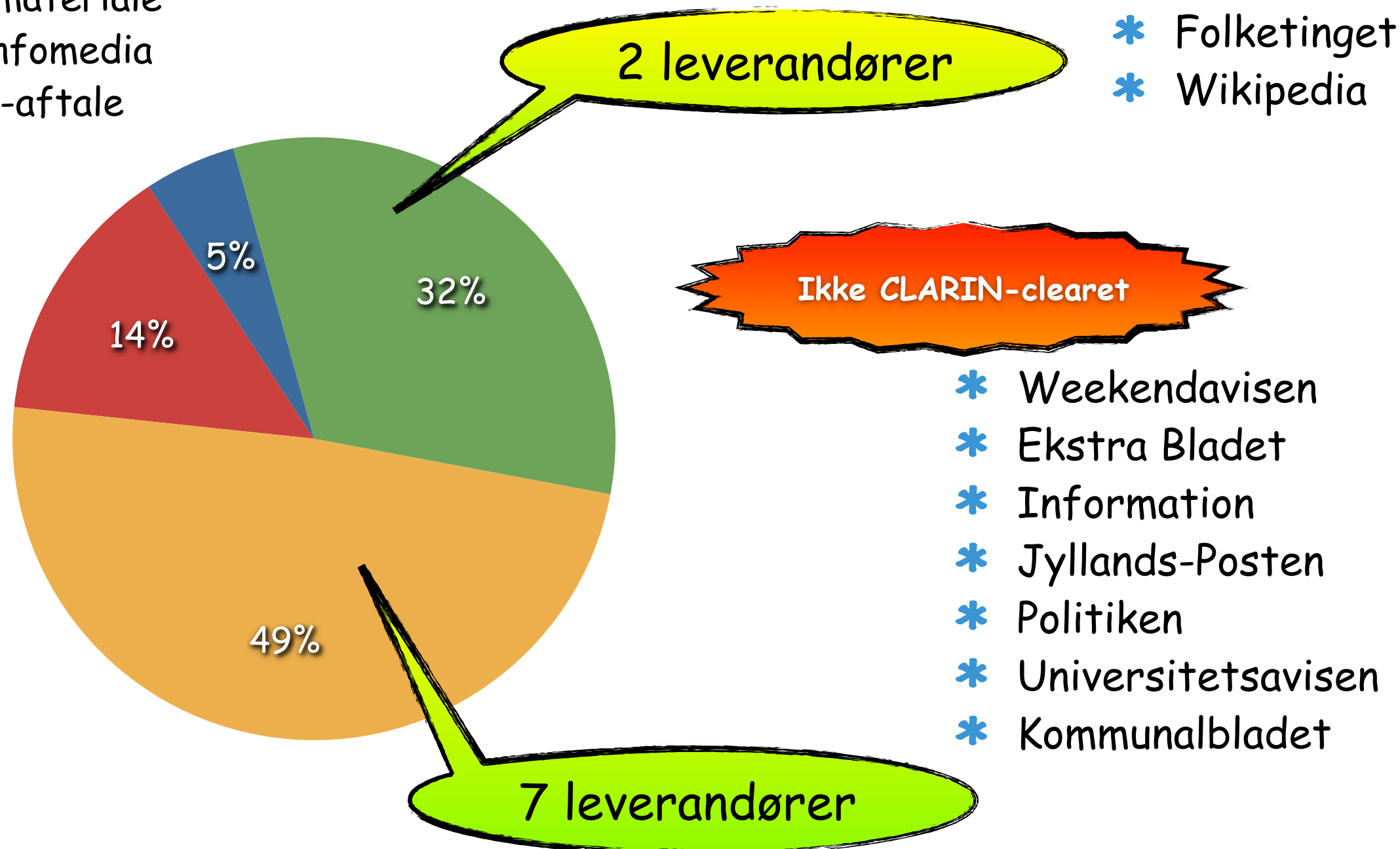
- * Via Infomedia (DSL): Helse, Magasinet Ejendom, Juristen, djøfbladet, Hus Forbi, Ældre Sagen, Højskolebladet, Folkeskolen, Samfundsøkonomen, penge.dk

- * Via webbet (DSN): blogsbjerg.com, skauogco.blogspot.com, jarlcordua.dk, Lifli, Bentes Blog, blogbogstaver.dk, selvhenter.org, Mangamania, clioonline.dk

Lad os se på en fordeling ...

1. Indsamling: Tekstmængde fordelt på aftaler

- CLARIN-aftale
- "Frit" materiale
- Lbd. Infomedia
- Fælles-aftale



Næste punkt: Metadata og formater ...

2. Metadata og formater: Kildebeskrivelsen

* Kildebeskrivelsen

- * Tekstittel
- * Forfatter
- * Oversætter
- * Værktitel
- * Udgiver
- * Forlag
- * Lokalisering
- * URL
- * Filnavn

```
<sourceDesc>
  <biblStruct>
    <analytic>
      <title xml:lang="languageId"
        level="titleLevel">textTitle</title>
      <author>
        <name ref="#personId">surname, forename</name>
      </author>
      <respStmt n="translators">
        <resp>Translated by</resp>
        <name ref="#personId">surname, forename</name>
      </respStmt>
    </analytic>
    <monogr>
      <title xml:lang="languageId">editionTitle</title>
      <editor>
        <name ref="#personId">surname, forename</name>
      </editor>
      <imprint>
        <publisher n="publId">publHouse</publisher>
        <date when="publDate" cert="certainty"/>
        <biblScope type="issue">edIssue</biblScope>
        <biblScope type="sect">edSect</biblScope>
        <biblScope type="vol">edVolume</biblScope>
        <biblScope type="chap">edChapter</biblScope>
        <biblScope type="pp">edPages</biblScope>
      </imprint>
    </monogr>
    <idno type="uri">textUri</idno>
    <idno type="file">textFileName</idno>
    <relatedItem type="relatedType">
      <bibl>
        <title xml:lang="languageId">relatedTitle</title>
        <idno type="ctb">relatedId</idno>
      </bibl>
    </relatedItem>
  </biblStruct>
</sourceDesc>
```

TEI P5

<SourceDesc>

Videre med tekstprofilen ...

2. Metadata og formater: Tekstprofilen

* Tekstprofilen

- * Tilblivelsesår
- * Sprogtype
- * Kanal
- * Forfatning
- * Afledning
- * Domæne
- * Faktualitet
- * Interaktion: Rolle
- * Interaktion: Alder
- * Forberedelse
- * Formål
- * Klassifikation

Se den fulde headerbeskrivelse under:
[http://korpus.dsl.dk/clarin/corpus-doc/
text-header.pdf](http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf)

```
<profileDesc>

  <creation>
    <date when="textCreationYear" cert="certainty" />
  </creation>

  <langUsage>
    <language ident="languageId"
      <u>languageCharacterization
    </language>
  </langUsage>

  <textDesc>
    <channel mode="tdChannelMode">tdChannel</channel>
    <constitution type="tdConstitutionType" />
    <derivation type="tdDerivationType">
      <lang>languageId</lang>
    </derivation>
    <domain type="tdDomainDiscourse">tdDomain</domain>
    <factuality type="tdFactualityType" />
    <interaction active="tdInteractActive"
      passive="tdInteractPassive"
      <note type="interactRole">tdInteractRole</note>
      <note type="interactAge">tdInteractAge</note>
    </interaction>
    <preparedness type="tdPrepType" />
    <purpose type="tdPurposeType" />
    <tdref scheme="myClassification" target="myValue" />
    <classCode scheme="theirClassification">theirValue</classCode>
  </textDesc>

  <textClass>
```

Videre med tekstprofilen ...

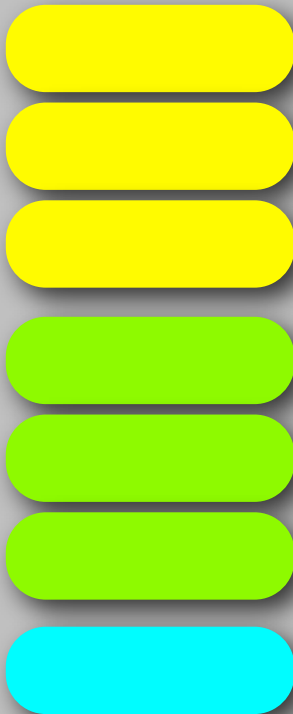
2. Metadata og formater: Tekstenhed

Hierarkisk format:
XML TEI P5

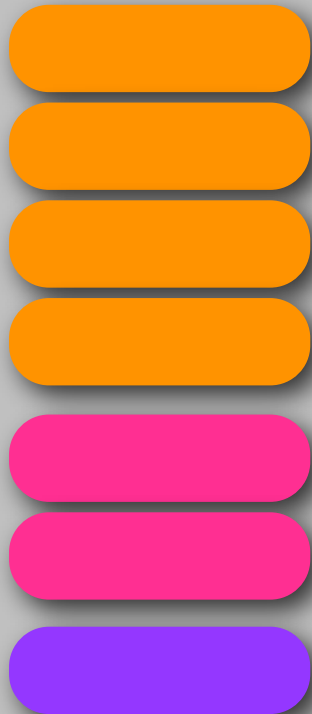
Tekstenhed

Header

Kildeinfo



Tekstprofil



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>
```

Intern
db-repræsentation?

Acceptabelt
tekstformat?

Videre med tekstformat ...

2. Metadata og formater: Tekstformat

Eksempel fra PAROLE-korpusset

Ej TEI P5

```
<p>
<s>
<W lemma="to" msd="AC--U--">To</W>
<W lemma="kendt" msd="ANP[CN]PU=[DI]U">kendte</W>
<W lemma="russisk" msd="ANP[CN]PU=[DI]U">russiske</W>
<W lemma="historiker" msd="NCCPU==I">historikere</W>
<W lemma="Andronik" msd="NP--U==--">Andronik</W>
<W lemma="Mirganjan" msd="NP--U==--">Mirganjan</W>
<W lemma="og" msd="CC">og</W>
<W lemma="Igor" msd="NP--U==--">Igor</W>
<W lemma="Klamkin" msd="NP--U==--">Klamkin</W>
<W lemma="tro" msd="VADR-----A-">tror</W>
<W lemma="ikke" msd="RGU">ikke</W>
<W lemma="," msd="XP">,</W>
<W lemma="at" msd="CS">at</W>
<W lemma="Rusland" msd="NP--U==--">Rusland</W>
<W lemma="kunne" msd="VADR-----A-">kan</W>
<W lemma="udvikle" msd="VAF-----P-">udvikles</W>
<W lemma="uden" msd="SP">uden</W>
<W lemma="en" msd="PI-CSU--U">en</W>
<W lemma=""" msd="XP">"</W>
<W lemma="jernnæve" msd="NCCSU==I">jernnæve</W>
<W lemma=""" msd="XP">"</W>
<W lemma="." msd="XP">.</W>
</s>
</p>
```

Annoteringer og tekst
filtret ind i hinanden

Interpunktion
fortolket som ord

Vanskeligt at tilføje
yderligere annoteringslag

Tokenreferencer ikke mulige

Annoteringer kan ikke
adskilles fra teksten

Alternative annoteringer af
samme tekst kræver ny version

Pga. disse problemer vælger vi en anden løsning ...

2. Metadata og formater: Tekstformat

DK-CLARIN tekstrepræsentation

Se den fulde tekstformatbeskrivelse under:
<http://korpus.dsl.dk/clarin/corpus-doc/text-format.pdf>

<code><w xml:id="x012">Mirganjan</w></code>	<code>NP--U==-</code>
<code><c xml:id="x013" type="s"/></code>	<code>CC</code>
<code><w xml:id="x014">og</w></code>	<code>NP--U==-</code>
<code><c xml:id="x015" type="s"/></code>	<code>NP--U==-</code>
<code><w xml:id="x016">Igor</w></code>	<code>VADR=----A-</code>
<code><c xml:id="x017" type="s"/></code>	<code>RGU</code>
<code><w xml:id="x018">Klamkin</w></code>	<code>XP</code>
<code><c xml:id="x019" type="s"/></code>	<code>CS</code>
<code><w xml:id="x020">tror</w></code>	<code>NP--U==-</code>
<code><c xml:id="x021" type="s"/></code>	<code>VADR=----A-</code>
<code><w xml:id="x022">ikke</w></code>	<code>VAF=-----P-</code>
<code><c xml:id="x023" type="p">,</c></code>	<code>SP</code>
<code><c xml:id="x024" type="s"/></code>	<code>PI-CSU--U</code>
<code><w xml:id="x025">at</w></code>	<code>XP</code>
<code><c xml:id="x026" type="s"/></code>	<code>NCCSU==I</code>
<code><w xml:id="x027">Rusland</w></code>	<code>XP</code>
<code><c xml:id="x028" type="s"/></code>	
<code><w xml:id="x029">kan</w></code>	
<code><c xml:id="x030" type="s"/></code>	
<code><w xml:id="x031">udvikles</w></code>	
<code><c xml:id="x032" type="s"/></code>	
<code><w xml:id="x033">uden</w></code>	
<code><c xml:id="x034" type="s"/></code>	
<code><w xml:id="x035">en</w></code>	
<code><c xml:id="x036" type="s"/></code>	
<code><c xml:id="x037" type="p">"</c></code>	
<code><w xml:id="x038">jernnæve</w></code>	
<code><c xml:id="x039" type="p">"</c></code>	

Næste punkt: Teksthåndtering

3. Teksthåndtering

* Tekstkonvertering

- * Indsamlede tekster udstyres med metadata
- * Tekster med metadata konverteres til internt format
- * Tekstenheder gemmes på et lager, i en tekstbank

* Tekstbanken

- * Opbevarer tekster i ensartet struktur
- * Giver adgang til forskellige operationer, fx ordklasseopmærkning
- * Grupper af tekster (korpora) udtrækkes herfra

Tekniske detaljer er beskrevet her:
[http://korpus.dsl.dk/clarin/corpus-doc/
text-processing.pdf](http://korpus.dsl.dk/clarin/corpus-doc/text-processing.pdf)

Hvor står vi nu?

Her står vi nu

1. Indsamling

2. Metadata og formater

Korpus 2010

3. Teksthåndtering

6. Adgang

4. Ordklasseopmærkning

5. Det færdige korpus

Infrastruktur

Lad os se nærmere på infrastrukturen ...

3. Teksthåndtering: Intro til korpusfabrikken

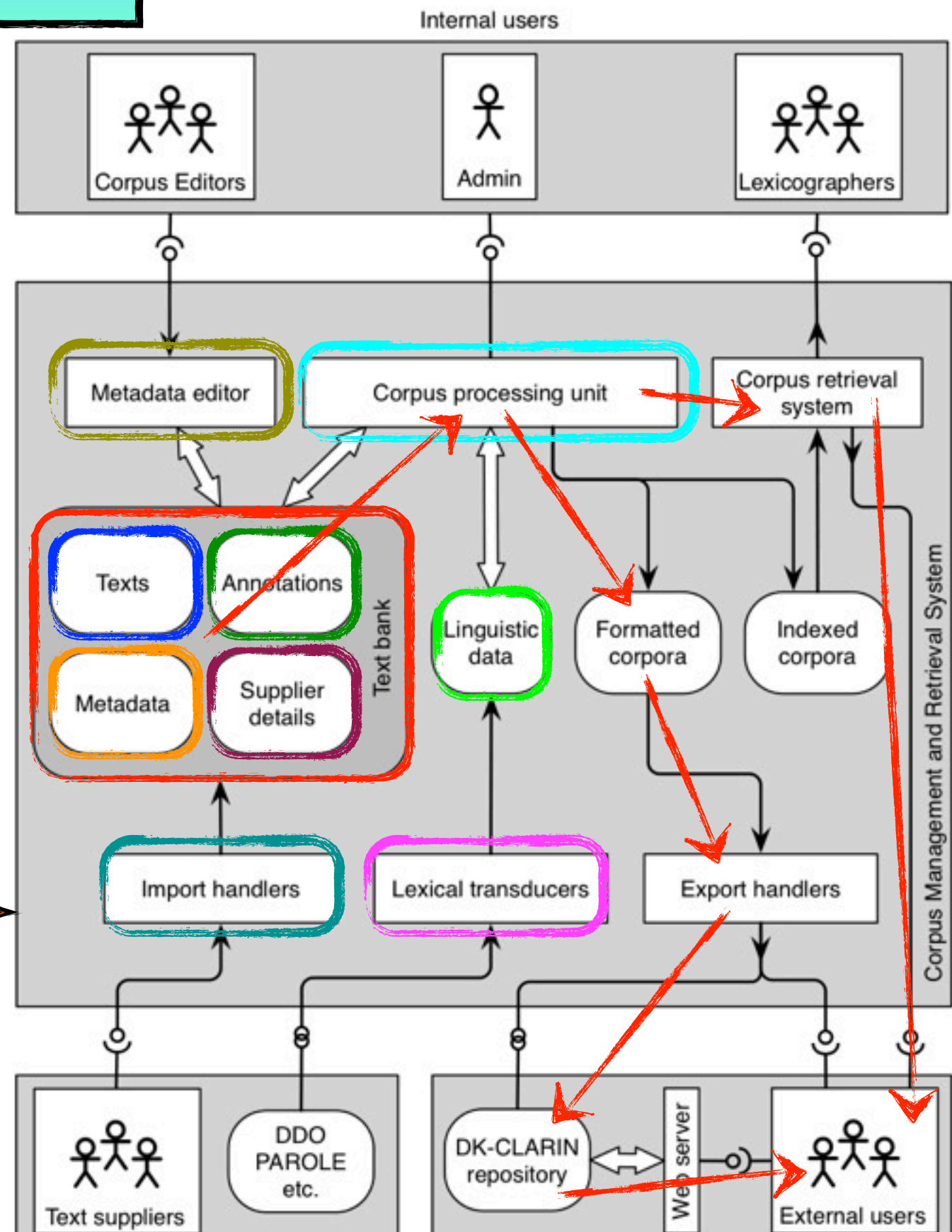
* Teksthåndtering

- * Tekstbanken er det centrale lager
- * Tekster, annoteringer og metadata holdes adskilt
- * Leverandøroplysninger
- * Indsamlede tekster importeres vha. særlige konverteringsprogrammer
- * Metadata bør kunnes redigeres
- * Annoteringer tilføjes via CPU'en
- * Hertil bruges sproglige data afledt af forskellige resurser

Der kan læses mere her:

<http://korpus.dsl.dk/clarin/corpus-doc/concepts.pdf>

Adgang?



Næste punkt: Ordklasseopmærkning

4. Ordklasseopmærkning: Valg af metode

* Kriterier

- * Åbenhed: Programmel & sprogviden
- * Tilgængelighed
- * Skal kunne lemmatisere
- * Udbredt programmeringssprog, helst Java
- * Skal kunne tilpasses forskellige behov
- * Veldokumenteret
- * Løbende vedligeholdelse

Læs evalueringsrapporten:
<http://korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf>

* Evaluering og konklusion

- * 12 taggere, heraf 2 til dansk
- * Feltforsøg med Sujit Pals HMM Tagger
- * Træningsgrundlag: PAROLE v.2
- * Fuldformsleksikon: DSL Flexion, DDO, PAROLE

Læs om taggeren:
<http://korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf>

Hvorfor ordklasseopmærkning?

4. Ordklasseopmærkning: Hvorfor?

Find form af verbet 'arbejde'
efterfulgt af præposition
efterfulgt af et substantiv

Formel søgning

[lemma="arbejde" & pos="V"]
[pos="PRP"]
[pos="N"]

- ▶ arbejder på højtryk (43)
- ▶ arbejder med børn (21)
- ▶ arbejder i døgndrift (17)
- ▶ arbejder på sagen (16)
- ▶ arbejde for føden (15)
- ▶ arbejder med gensplejsning (15)
- ▶ arbejder på deltid (15)
- ▶ arbejde med mennesker (14)
- ▶ arbejder i udlandet (14)
- ▶ arbejder med planer (14)
- ▶ arbejdet på højtryk (13)
- ▶ arbejde i grupper (12)
- ▶ arbejde med børn (12)
- ▶ arbejder med sagen (12)
- ▶ arbejder for øjeblikket (11)
- ▶ arbejder om natten (11)
- ▶ arbejdet med sagen (11)
- ▶ arbejde i udlandet (10)
- ▶ arbejde på deltid (10)
- ▶ arbejde på højtryk (10)
- ▶ arbejdede i døgndrift (10)
- ▶ arbejdede på højtryk (10)
- ▶ arbejde om aftenen (9)
- ▶ arbejdede for tyskerne (8)
- ▶ arbejder med mennesker (8)
- ▶ arbejder på kontor (8)
- ▶ arbejdet på sagen (8)
- ▶ arbejde i haven (7)

Eller som konkordans ...

4. Ordklasseopmærkning: Konkordans

sur. Konservative danskere " Laboremus pro patria "- **arbejd for fædrelandet**- har den gamle brygger Jacobsen ladet skrive over indgangsdøren. Men
ADJ ADJ N USPEC USPEC USPEC V PRP N V ART ADJ N PROP V V PRP N KC

noget gjort ved, ellers kommer du et' langt med æ **arbejd' i daw.** Ejnersterede og mente, det nok skulle komme til at
INDP V PRP ADV V PERS ART ADV ADV IN V PRP N PROP KC MEM

ovenpå og nedenunder. Husk af og til at bytte partnere. **Arbejd i grupper** på fire
ADV KC ADV ADV ADV ADV ADV INFM V N V PRP N PRP

der siges, skal passe til bevægelserne. Ingen må styre!! **Arbejd i grupper** på
INDP V V V PRP N INDP V V V PRP N PRP

til den første og sagde: Min søn, gå ud og **arbejd i vingården** i
PRP ART NUM KC V DET N V ADV KC V PRP N ADV ADV

i MAD FOR DIG 1 og 2 retter fra udlandet. **Arbejd med emnet:** "Fremmed "mad " i flere hjemkundskabstimer og prøv så mange forskellige
PRP PROP PROP PROP PROP KC NUM N PRP N V PRP N USPEC USPEC PRP DET N KC V ADV DET ADJ

[...] Derfor, mine kære, I, som altid har været lydige: **Arbejd med frygt** og bæven på jeres frelse, ikke blot som da jeg
ADV DET ADJ PERS INDP ADV V V ADJ V PRP N KC N PRP PERS N ADV ADV PRP KS PERS

hårdt mod bogens ryg, så du til sidst knækker den. **Arbejd med hænderne** fra side til side, så der kommer en revne på
ADV PRP N N KS PERS ADV ADV V PERS V PRP N PRP N PRP N KS ADV V ART N PRP

klog af at studere. Men først og sidst: Øv dig. **Arbejd under kameraet!** Det kan godt være, det ikke bliver så godt i
ADJ PRP INFM V KC ADV KC ADV V PERS V PRP N PERS V ADV V PERS ADV V ADV ADJ PRP

Samtidig er der regler for, hvor længe børnene må **arbejde ad gangen**, når de er under 15 år- nemlig højst to timer
ADV V ADV N PRP ADV ADV V V PRP N KS PERS V ADV NUM N ADV ADV NUM N

mindre beløb. Og det kom de to unger til at **arbejde af. -Drengene** blev ansat i det firma, der skulle have penge tilbage
DET NUM PRP INFM V PRP N V V PRP DET N INDP V V N ADV

videnskabeligt **arbejde af videns** og ideologiformidlende og administrativ karakter. Størstedelen af en alme
ADV V PRP N KC V KC ADJ N N PRP ART

i D. Hilden skal **arbejde bag kulisserne** i redaktionen af debatprogrammet, der fra september bl.a. vil bet
PRP PRP N V V PRP N PRP N PRP N INDP PRP N ADV V

politikere uden store armbævere. Han kan bedst lide at **arbejde bag linierne** på sin egen stille facon og er ikke vild med
N PRP ADJ N PERS V ADV V INFM V PRP N PRP DET DET ADJ N KC V ADV ADJ PRP

at have de fleste af de 150-200 mennesker, der skal **arbejde bag scenen**, men er altid interesseret i ny arbejdskraft. Folk, der kunne
INFM V ART DET PRP ART NUM N INDP V V PRP N KC V ADV ADJ PRP ADJ N N INDP V

Grev Olaf. "En køn pige som du skal sandelig ikke **arbejde bag scenen.** " "Jamen jeg vil gerne ", sagde Violet. Grev Olafs ene øjenbryn
PROP PROP USPEC ADJ N INDP PERS V ADV ADV V PRP N USPEC PERS V USPEC V PROP PROP PROP ADJ N

Og så prøver vi det samme
i Google ...

Eller
Infomedia ...

Næste punkt: Det færdige korpus ...

5. Det færdige korpus

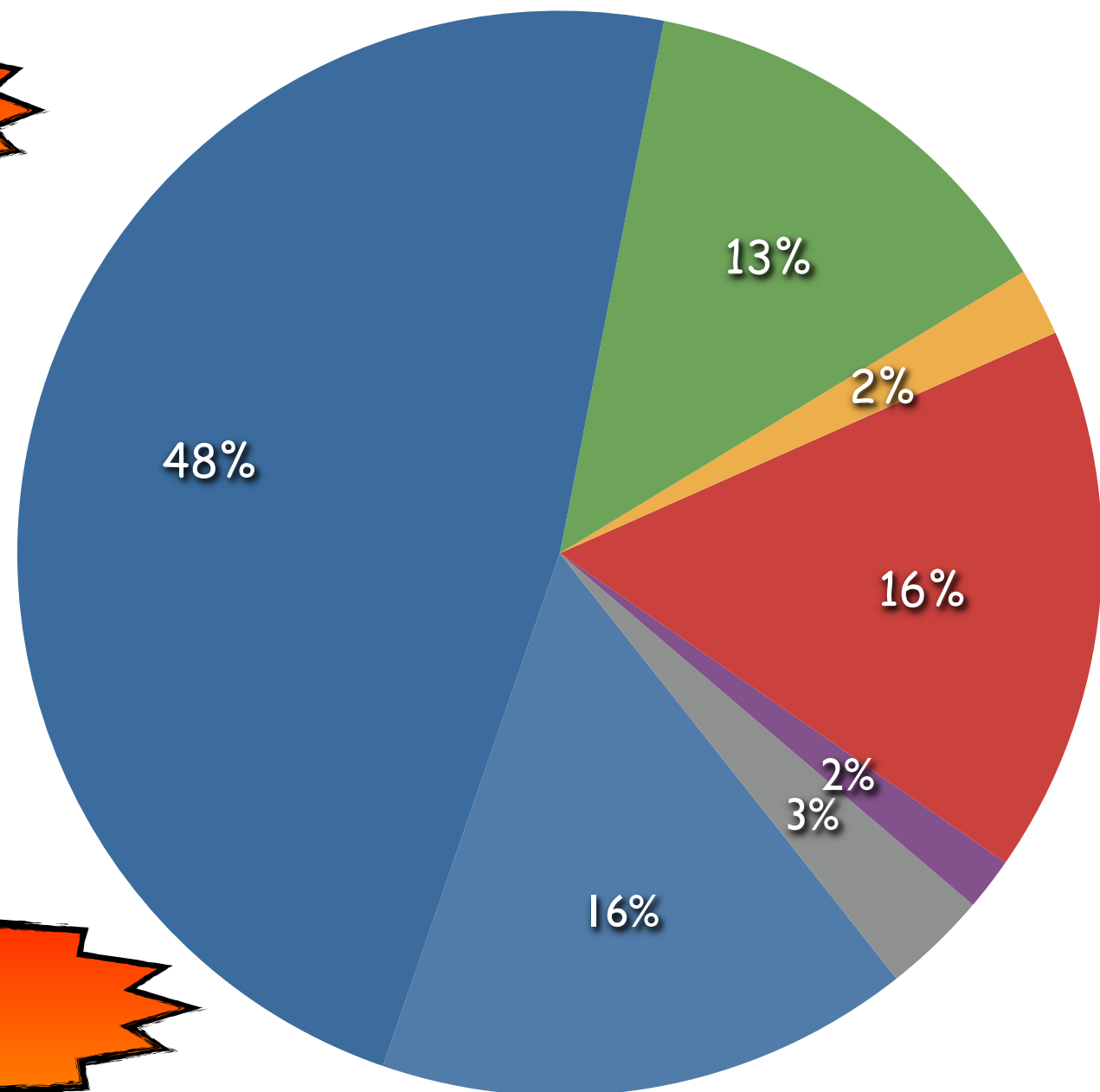
45 mio. ord

2008-2011

- avis
- blad
- blog
- folketing
- forum
- web
- wikipedia

Lager på >300 mio.

Fordeling på teksttyper



Sidste punkt: Adgang ...

6. Adgang

Korpus findes som XML
TEI P5-filer

POS-tagget
&
lemmatiseret

Uploades p.t. til
DK-CLARIN repositoret

Gøres p.t. søgbart i særlig
web-grænseflade

Hvad der DK-CLARIN-repositoriet?

6. Adgang: CLARIN-repositoriet



TAK!