

Hvad?

Korpus 2010 & ePOS

Det almensproglige korpus i
DK-CLARIN

En ny POS-tagger
for dansk

Hvem?

Jørg Asmussen
Det Danske Sprog- og Litteraturselskab

Hvornår & hvor?

Onsdag, 31. oktober 2012
Sprogteknologisk workshop, KU

Hvad ellers?

Denne præsentation findes under
<http://korpus.dsl.dk/staff/ja/pres/sprogtekno2012.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.

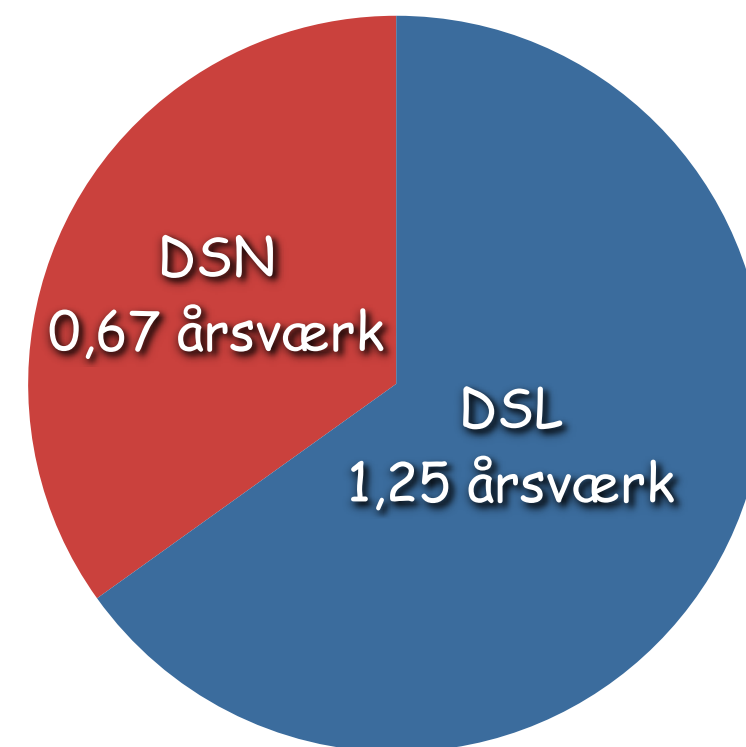
Flere oplysninger:
<http://clarin.dsl.dk>

* Deltagere:

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

* 1 million kroner:

- * DSL 700.000
- * DSN 300.000



Arbejdsopgaver ...

Arbejdsopgaver = Program

1. Indsamling

2. Metadata og formater

Korpus 2010

5. Adgang

3. Ordklasseopmærkning

4. Det færdige korpus

Indsamling ...

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:

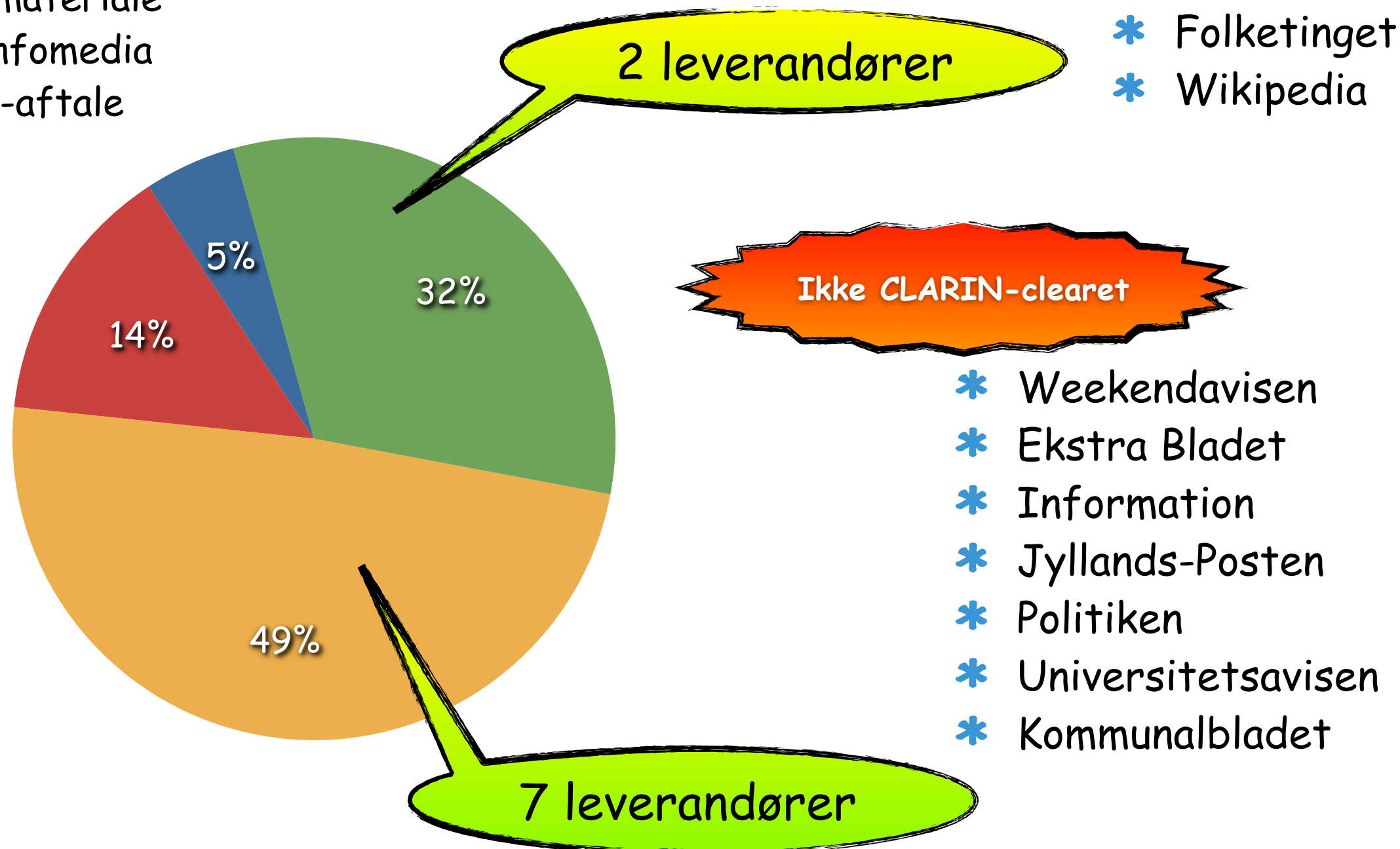
- * Via Infomedia (DSL): Helse, Magasinet Ejendom, Juristen, djøfbladet, Hus Forbi, Ældre Sagen, Højskolebladet, Folkeskolen, Samfundsøkonomen, penge.dk

- * Via webbet (DSN): blogsbjerg.com, skauogco.blogspot.com, jarlcordua.dk, Lifli, Bentes Blog, blogbogstaver.dk, selvhenter.org, Mangamania, clionline.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</u>
    </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: 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="&quot;" msd="XP">"</W>  
<W lemma="jernnæve" msd="NCCSU==I">jernnæve</W>  
<W lemma="&quot;" 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: Ordklasseopmærkning

3. Ordklasseopmærkning: Muligheder

* 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](http://korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf)

* Evaluering og konklusion

- * 12 taggere, heraf 2 til dansk
- * Feltforsøg med en Java-implementering af en HMM Tagger
- * Træningsgrundlag: PAROLE v.2
- * Fuldformsleksikon: DSL Flexion, DDO, PAROLE

Læs om taggeren:

[http://korpus.dsl.dk/clarin/corpus-doc/
pos-design.pdf](http://korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf)

Fremgangsmåde...

3. Ordklasseopmærkning: Fremgangsmåde

- * **Princip: Så simpelt som muligt**
 - * Ingen lingvistisk viden nødvendig til tekstsegmentering, dvs. intet lingvistisk sætnings- eller ordkoncept
- * **Tilpasning af PAROLE-korpusset (ca. 9000 ændringer)**
 - * Ordgrænser: mellemrum, interpunktion, bindestreger
 - * Ortografiske fejl og taggingfejl blev rettet
- * **Opbygning af passende fuldformsleksikon (ca. 100000 lemmaer)**
 - * Bygger på diverse DSL-resurser, herunder DDO
 - * Indtil videre råversion, kræver manuel efterbehandling
- * **Træningsmetode**
 - * Slå ordform op i leksikonet
 - * Transformér ordformen til et mønster af potentielle tags
 - * Brug dette mønster + det rigtige tag som input i træningen

Tag-struktur...

3. Ordklasseopmærkning: Tag-struktur

- * Princip: Strengt positionel (ikke kompakt)
- * Find alle ord med en genitiv-markør uanset deres ordklasse!

XY:abcd:ef:ghij

Klassifikation

X = ordklasse

Y = subklasse

Nominale træk

a = numerus

b = definitthed

c = kasus

d = genus

Verbale træk

e = tempus

f = diatese

Diverse træk

g = komparation

h = person

i = reflektivitet

j = possessor

Læs om tag-strukturen:

[http://korpus.dsl.dk/clarin/corpus-doc/
pos-design.pdf](http://korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf)

Eksempel...

3. Ordklasseopmærkning: Eksempel

<s>			
Min	—	min	PO:s--c:--:-1ns
far	—	far	NC:siuc:--:-----
pacede	—	pace	VF:-----:ta:-----
mig	—	jeg	PP:s-uc:--:-1#-
frem	,_	frem	D-:-----:--:u---
men	—	men	CC:-----:--:-----
jeg	—	jeg	PP:s-nc:--:-1n-
måtte	—	måtte	VF:-----:ta:-----
stoppe	—	stoppe	VI:-----:-a:-----
som	—	som	US:-----:--:-----
17	-	17	LW:-----:--:-----
årig	,_	årig	AC:siuc:--:p---
fordi	—	fordi	CS:-----:--:-----
jeg	—	jeg	PP:s-nc:--:-1n-
begyndte	—	begynde	VF:-----:ta:-----
at	—	at	UI:-----:--:-----
tage	—	tage	VI:-----:-a:-----
speed	.\$	speed	NC:siuc:--:-----
</s>			

ID	2009016415
Title	Minderne driver mig på gaden
Creator	Høegh, Birgitte Ellemann (author)
Edition	Hus Forbi
Publisher	Hus Forbi
Date	2010-12-02
Location	1; 32

Det færdige korpus...

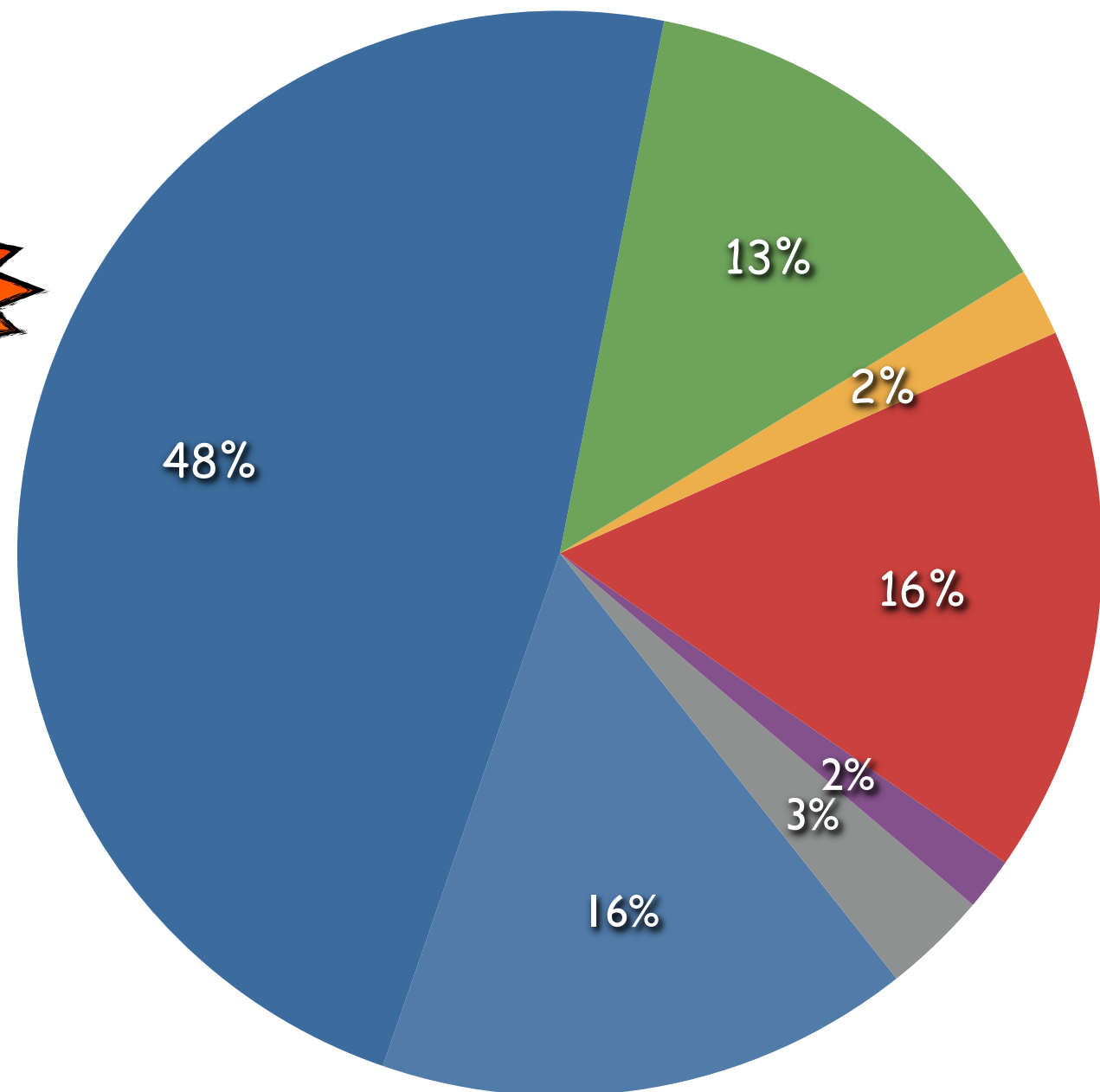
4. Det færdige korpus

45 mio. ord

2008-2011

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

Fordeling på teksttyper



Sidste punkt: Adgang ...

5. Adgang

Korpus findes som XML
TEI P5-filer

POS-tagget
&
lemmatiseret

Uploade[ts] til
DK-CLARIN repositoret

Bliver søgbart i særlig
web-grænseflade

TAK!