CLARINO

Common Language Resources and Technology Infrastructure – Norway

Koenraad De Smedt

Universitas Bergensis

OSLO · DIE XVI FEBRUARII A·D· MMXII

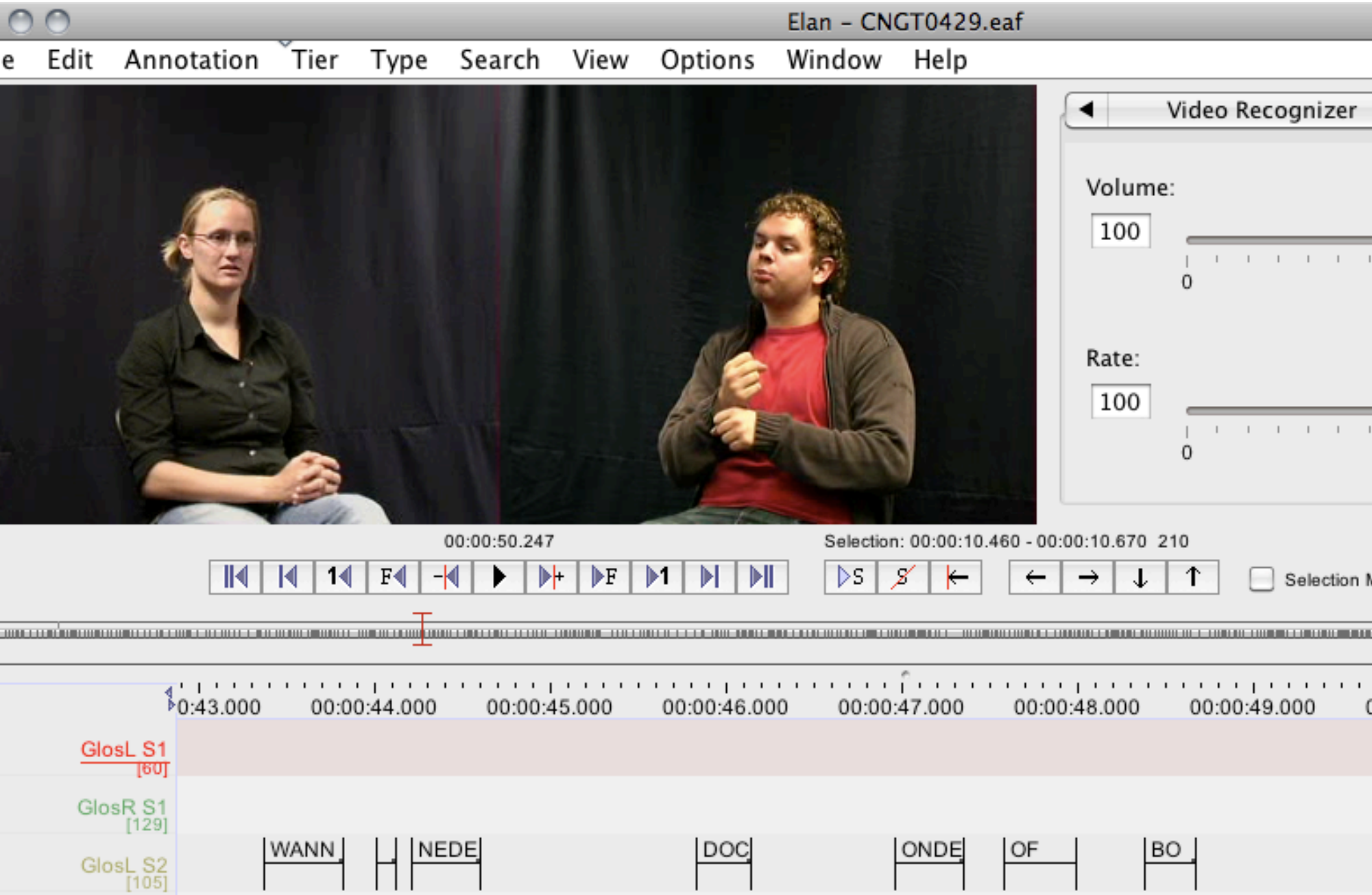
Hvorfor er språkdata viktig?

- Nesten all menneskelig kommunikasjon er i språklig form: aviser, bøker, blogger, sms, tv, samtaler osv.
- Mange samlinger av kunnskap er i språklig form: biblioteker, arkiver, kirkebøker, osv.
- Sekundære språkdata og verktøy: ordbøker, nøkkelord, grammatikker, frekvenslister, osv.
- Språkverktøy: maskinoversettelse, automatisk sammendrag, indeksering, osv.
- Eksperimentelle data: øyebevegelser, reaksjonstider, spektogrammer, EEG, fMRI, osv.

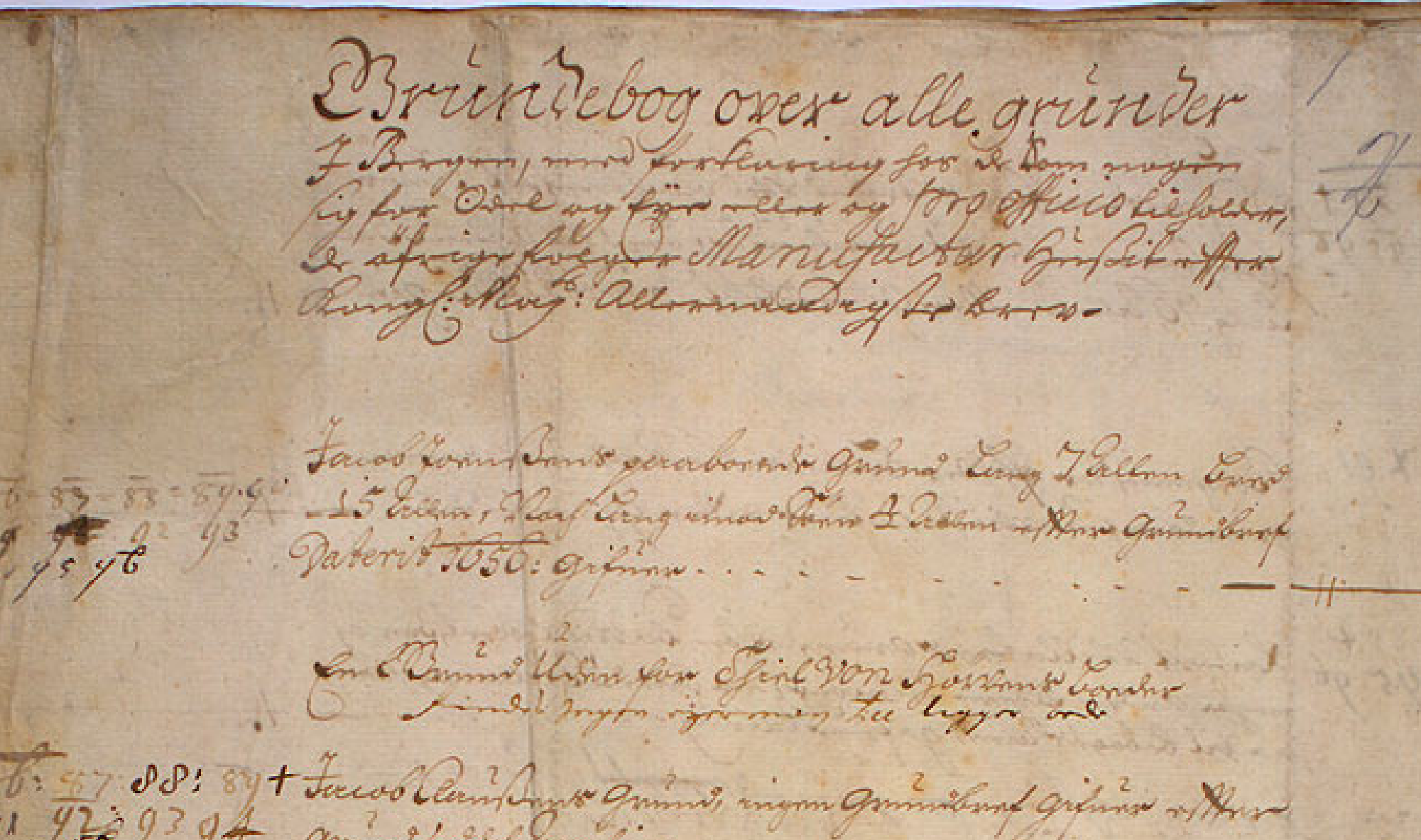
Hvorfor er språkdata viktig?

Språk er vår fremste kilde til informasjon om mennesker, kulturer, ideer og kognisjon.

Språkdata: eksempler



Språkdata: eksempler



Språkdata: eksempler

Grundbog over alle grunder

j Bergen, med forklaring hos de som nogen

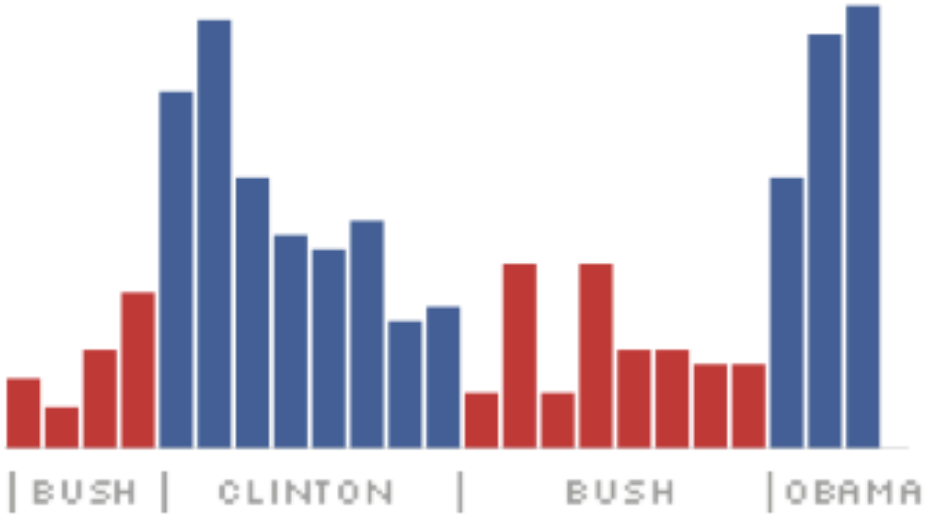
sig for odel og eye eller og pro officio tilholder,

de øfrige følger Manufactur hußit effter

Kongl: Mayts: allernaadigste brev.

		[Nykirkesokn]
1-5	bet. 86, 87, 88, 89, 90, 91. 93, 93, 94, 95, 96	Jacob Joenßens paaboende grund, *lang 7 allen, bred 15 allen,* noch lang imod søen *4* allen, effter grundbref daterit 1656. Gifuer
		En grund uden for Thiel von Høwens boeder. Findes jngen eyermand till ligger oede
1-1	bet 86, 87, 88, 89 + 91, 92, 93, 94, 95, 96	Jacob Claußens grund, ingen grundbref, gifuer effter grundseddel aarlig
1-1	bet 86, 87, 88, 89, 90, 91, 92, 93, 94, 95	Jacob Claußens grund uden for Thiel von Høwens boeder, *bred 28 allen,* og lang saa vit bebygges kand effter

Språkdata: eksempler



"Jobs" in State of the Union speeches

Språkdata: eksempler

Event type: **Arrest**
Who is injured: **Thai civilians ; an Iranian**
Number injured: **5**
Who is arrested: **Iranian suspects**
Number arrested: **2**
Weapons: **bomb**
Snippet: **Iranian bomb suspects arrested in Bangkok ...**
Place: **Bangkok.** *i*
Geo Path: Bangkok:Thailand lat: 13.7703017743177 lon: 100.626440699347

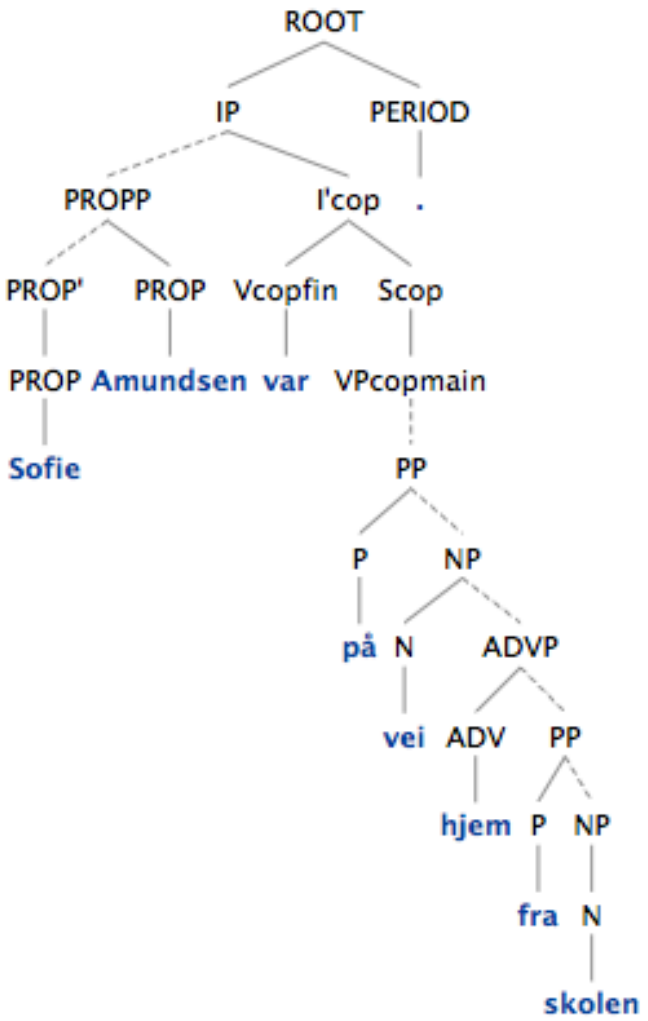
Israel: Thai bombs similar to those in India blast (AP)
Wednesday, February 15, 2012 12:07:00 PM CET
AP - The homemade "sticky" bombs discovered in a Bangkok house after an accidental blast were similar to devices used against Israeli Embassy targets in India and Georgia, Israel's ambassador said Wednesday, building on his country's claims the incidents are part of a covert terror campaign by Iran....

More articles

Events detection, JRC EMM NewsExplorer

Språkdata: eksempler

C-structure



F-structure

```
PRED      'være<[30:Amundsen], [19:på]>'
TNS-ASP   40 TENSE past, MOOD indicative
TOPIC     PRED      'Amundsen'
          NTYPE     NSEM 38 PROPER 39 PROPER-TYPE name
                    37 NSYN proper
          NAME-MOD  { PRED  'Sofie'
                      NTYPE NSEM 35 PROPER 36 PROPER-TYPE
                            34 NSYN proper
                      GEND  33 NEUT -, MASC -, FEM +
                    31 32 REF +, PERS 3, NUM sg, DEF +
          30 REF +, PERS 3, NUM sg, DEF +, CASE nom
PREDLINK  PRED 'på<[24:vei]>'
          OBJ       PRED      'vei'
                    NTYPE     NSEM 29 COMMON count
                              28 NSYN common
                    GEND      27 NEUT -, MASC +, FEM -
                              PRED      'hjem'
                                        PRED 'fra<[9:skole]>'
                                                  PRED  'skole'
                                                  NTYPE NSE
```

Hvorfor CLARIN?

Det finnes mange språkrelaterte datasamlinger og verktøy, men de er ofte usynlige, ikke katalogisert, ikke gjenbrukbare, ikke koblet til andre ressurser, ikke tilstrekkelig dokumentert, ikke alltid tilgjengelige, ikke kompatible med dataplattformer, ikke i riktige formater, ikke i samsvar med gjeldende standarder, osv.

Hvis bare en brøkdel av disse problemene kan løses, vil det være et stort fremskritt.

CLARINs mål

“CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities.”

Konsolidering, samling, tilgjengeliggjøring, analyse og utnyttelse av språkressurser som ellers kan gå tapt.

CLARINs visjon

- En forsker logger inn ved sin egen institusjon (f.eks. via Feide) og kommer inn på CLARIN-portalen.
- Hun ser i en katalog, søker i metadata eller søker i innhold.
- Hun lager sin egen virtuelle samling av ressurser fra flere databaser i inn- og utland.
- Hun signerer lisensavtaler der nødvendig.
- Hun spesifiserer en arbeidsflyt for prosessering av den virtuelle samlingen ved hjelp av analyseverktøy gjennom webtjenester og fjernbruk av tungregning.
- Resultatene blir lagret i et eget arbeidsområde.
- Data og metadata lastes opp i en database og den virtuelle samlingen lagres for fremtidig bruk ved hjelp av PIDer.

CLARIN: STATUS

- Prosjekt på ESFRI-veikartet
- 26 europeiske land er med
- Forberedende fase fra jan. 2008 til slutten av juni 2011
- ERIC rettsperson i 2012
- Implementering gjennom nasjonale prosjekter — CLARINO

CLARIN: LAG

1. Koordinasjon og styring (ERIC, europeisk konsortium)
2. Infrastruktur (langsiktig nasjonalt ansvar) — CLARINO
3. Innhold inkl. tilrettelegging (kortsiktige prosjekter støttet av Forskningsrådet, institusjonene mm.)

Nasjonalt nettverk av noder

Type A. Nasjonalbiblioteket, Uninett

Type B. Tekstlaboratoriet (UiO); EDD (UiO); LAP/IFI (UiO); UiB+NHH+UniResearch

Type C. NTNU, UIT

Ansvarsfordeling for tjenester

- *Infrastrukturtjenester* (Nasjonalbiblioteket, Uninett). Nasjonal katalog over språkressurser, langtidslagring, HPC-tilgang, autentisering og autorisering, PID-tjeneste.
- *Språkdatatjenester* (UiO, UiB, Uni Research, NHH). Korpusanalyse, terminologiportal, elektronisk utgaveplattform.
- *Språkteknologitjenester* (UiO, UiB, Uni Research). Språkanalyseportal (LAP), verktøy og prossesseringskjeder.
- *Levering av data og metadata* (alle noder).

Forhold til bibliotek

"Language infrastructures represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architecture"

Institusjonelle og nasjonale strategier ønskes mht:

- langsiktig arkivering av data
- kobling av publikasjoner til data
  data $\rightarrow$ publikasjon $\rightarrow$ data
- uttelling for publisering av data
- IPR, lisensiering av data til gjenbruk

Bibliotek som eHumanities-miljø: Perseus

Plato, *Republic*

Search

("Agamemnon", "Hom. Od. 9.1", "denarius")

All Search Options [view abbreviations]

Home | Collections/Texts | Research | Grants | Open Source | About | Help

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position. Hide browse bar

book:
section:

This text is part of:

Greek and Roman Materials
Greek Prose
Greek Texts
Plato
Plato, Republic

View text chunked by:

book : page
book : section
page
section

Table of Contents:

▼book 1
section 327a
section 327b
section 327c
section 328a
section 328b

Plat. Rep. 1.327a

Click on a word to bring up parses, dictionary entries, and frequency statistics

[327α]

Σωκράτης

κατέβην χθὲς εἰς Πειραιᾶ μετὰ Γλαύκωνος τοῦ Ἀρίστωνος προσευξόμενός τε τῇ θεῷ καὶ ἅμα τὴν ἑορτὴν βουλόμενος θεάσασθαι τίνα τρόπον ποιήσουσιν ἅτε νῦν πρῶτον ἄγοντες. καλὴ μὲν οὖν μοι καὶ ἡ τῶν ἐπιχωρίων πομπὴ ἔδοξεν εἶναι, οὐ μέντοι ἧττον ἐφαίνετο πρέπειν ἣν οἱ Θρᾷκες ἔπεμπον.

Plato. Platonis Opera, ed. John Burnet. Oxford University Press. 1903.

The Annenberg CPB/Project provided support for entering this text.

XML

An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.

Notes (James Adam) focus hide

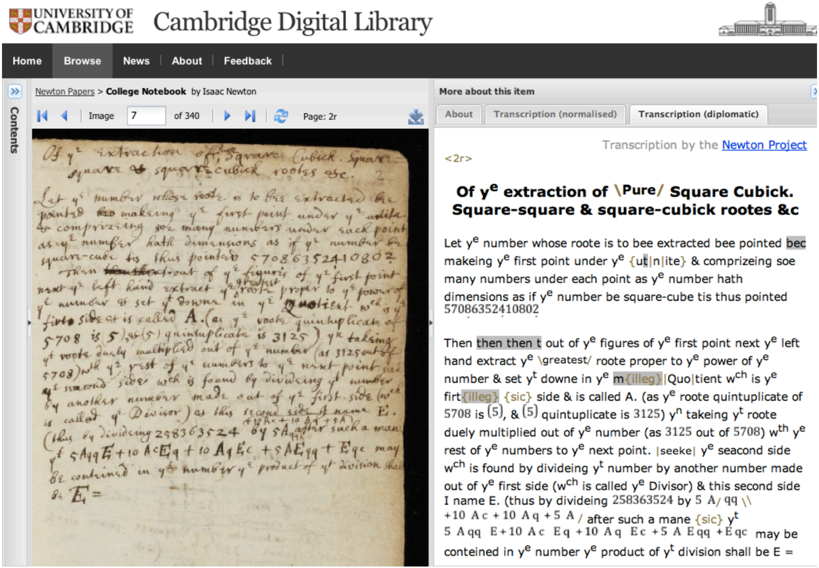
327A - 328B Socrates describes how he visited the Piraeus in company with Glauco, and was induced by Polemarchus and others to defer his return to Athens.

κατέβην κτλ. Dionys. Hal. de comp. verb. p. 208 (Reiske) ὁ δὲ Πλάτων, τοὺς ἑαυτοῦ διαλόγους κτενίζων καὶ βοστρυχίζων, καὶ πάντα τρόπον ἀναπλέκων, οὐ διέλιπεν ὀγδοήκοντα γεγονὼς ἔτη. πᾶσι γὰρ δή που τοῖς φιλολόγοις γνώριμα τὰ περὶ τῆς φιλοπονίας τάνδρὸς ἱστορούμενα, τά τ' ἄλλα, καὶ δὴ καὶ τὰ περὶ τὴν δέλτον ἣν τελευτήσαντος αὐτοῦ λέγουσιν εὑρεθῆναι ποικίλως μετακειμένην τὴν ἀρχὴν τῆς πολιτείας ἔχουσαν τήνδε "κατέβην χθὲς εἰς Πειραιᾶ μετὰ Γλαύκωνος τοῦ Ἀρίστωνος." See also Quint. VIII 6. 64, and Diog. Laert. III 37. The latter gives as his authorities Euphorion and Panaetius. As Cicero was tolerably familiar with the writings of Panaetius, it is possible that he too has the same story in view in de Sen. V 13, where he says of Plato ""scribens est mortuus."" The anecdote may well be true, but does not of course justify any inference as to the date of composition of the Republic. See Introd. § 4.

τῇ θεῷ. What goddess? Bendis or Athena? The festival is the Bendideia (354 A) and it is perhaps safest to acquiesce in the usual view that Bendis is here meant. "Alii Minervam intelligunt, quae vulgo ἡ θεὸς appellabatur; neque mihi videtur Socrates in ista Panathenaeorum propinquitate de Minerva veneranda cogitare non potuisse: sed quod simpliciter τὴν ἑορτὴν dicit, numina diversa statuere non sinit" (Schneider). We hear of a temple of Bendis in the Piraeus in 403 B.C. (τὴν ὁδὸν ἡ φέρει πρός τε τὸ ἱερὸν τῆς Μουνυχίας Ἀρτεμίδος καὶ τὸ Βενδίδειον **Xen. Hell. II 4. 11**). See also Introd. § 3 and App. I.

νῦν πρῶτον. Perhaps 410 B.C. Introd. § 3.

Bibliotek som eHumanities-miljø: Cambridge



CLARIN-ES

Inicio | Herramientas | Recursos | Documentación | Catalogo | Contacto

Category	Service name	Description
Chunking segmentation		
	iula_preprocess (WSDL)	***cat**: Preprocés de textos (el servei de preprocés requereix que el text d'entrada estigui en format text pla (file.txt) i UT Bàsicament, el preprocés s'encarrega de (i) segmentar el text en unitats estructurals menors (títols, paràgrafs, oracion (ii) detectar entitats que no es troben als diccionaris (nombres, abreviatures, URLs, correus electrònics, noms propis, e mantenir en un únic bloc seqüències de dos o més mots (dates, locucions, noms propis, etc.).* ***es**: Preproceso de textos (el servicio de preproceso requiere que el texto de entrada esté en formato de texto plano (f en UTF-8. Esencialmente, el preproceso se encarga de: (i) segmentar el texto en unidades estructurales menores (títul párrafos, oraciones, etc.); (ii) detectar entidades que no se encuentren en los diccionarios (números, abreviaturas, URL electrónicos, nombres propios, etc.); y (iii) mantener en un único bloque secuencias de dos o más palabras (fechas, loc nombres propios, etc.).* ***en**: Text preprocess. (this preprocess service requires that the input text be in plain text format (file .txt) and UTF-8. B carries out: (i) text segmentation into minor structural units (titles, paragraphs, sentences, etc.); (ii) detection of entit found in dictionaries (numbers, abbreviations, URLs, emails, proper nouns, etc.); and (iii) the keeping of sequences of t more words in a single block (dates, phrases, proper nouns, etc.).*

Run service

COMPLETED

Result	Size	Type
output	565	text
report	919	text
detailed_status	1	unknown

Remove

Inputs

input

as URL

La canciller alemana, Angela Merkel, y el presidente francés, Nicolas Sarkozy, han asegurado hoy que las negociaciones del pacto fiscal en la eurozona concluirán en las "próximas semanas". Quieren que los acuerdos bilaterales se firmen, como

direct data or local file

Choose File no file selected

language es

mandatory

Report

```
Summary:
        Completed: Maybe
        Termination status: 0
        Started:   2012-ene-09 17:33:29 (CET)
        Ended:     2012-ene-09 17:33:38 (CET)
        Duration:  0:00:09.531
Report:
Some error messages were reported.

Name: chunking_segmentation.iula_preprocess
Job ID: [chunking_segmentation.iula_preprocess]5b8df9d1.1346af4b51d._7fed
Program and parameters:
/usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hec
-inputtext
i_input
-language
es
```

Forhold til Språkbanken

Språkbanken:

- *"det naturlige samlingspunktet for lagring og distribusjon av offentlige og private digitale språkressurser"*
- *"å være infrastruktur i språkteknologisk forskning, utvikling og produkt- og tjenestetilpasning for norsk språk"*

CLARIN

- Europeisk satsing som inkluderer alle språk
- Alle slags språkrelaterte forskningsbehov innen humanistiske fag (f.eks. Wittgensteinarkivet; Bosnisk-korpuset)
- Språkteknologi er til dels et instrument heller enn et mål

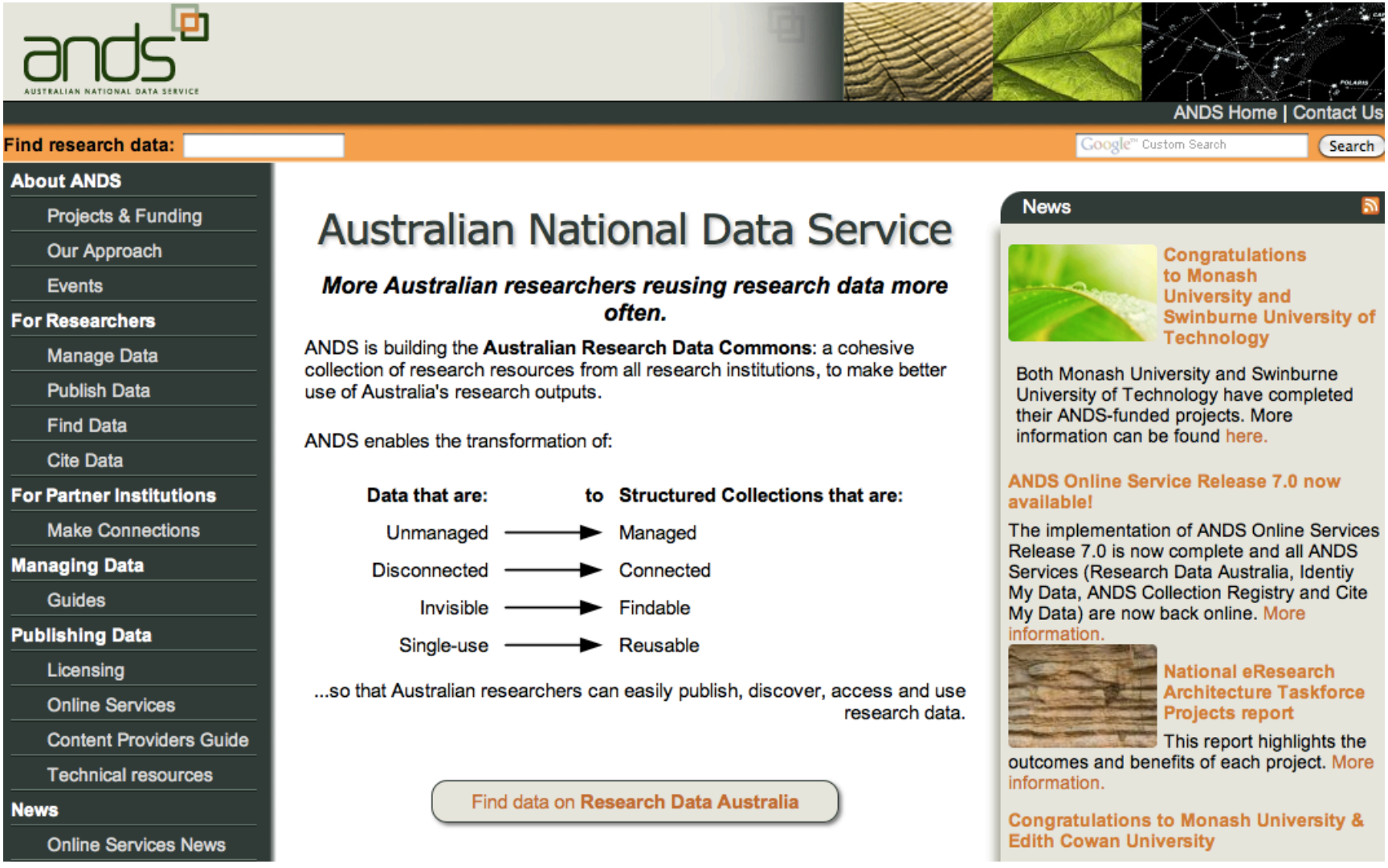
DASISH

Europeisk INFRA 2011 prosjekt, samarbeid mellom:

- CESSDA (Council of European Social Science Data Archives)
- CLARIN (Common Language Resources and Technologies Infrastructure)
- DARIAH (Digital Research Infrastructure for the Arts and Humanities)
- ESS (European Social Survey)
- SHARE (Survey of Health, Ageing and Retirement in Europe)

Felles strategier for forskningsdata i samfunnsvitenskap og humanistiske fag

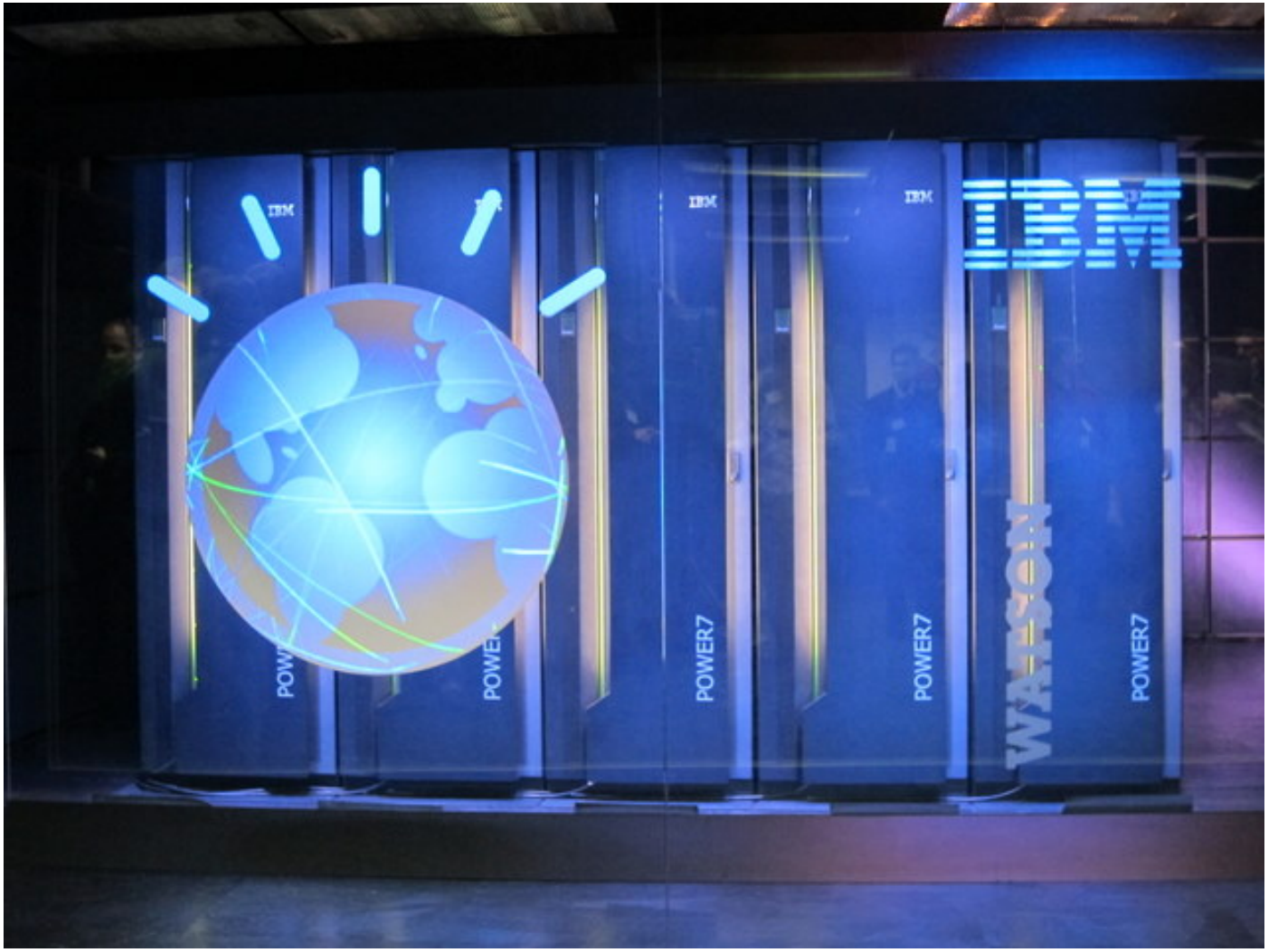
Mulig driftsmodell: ANDS



Forhold til Notur/NorStore

- Lagring og tungregning i CLARINO (også for relatert infrastruktur f.eks. INESS)
- Trenger mer permanent strategi for eInfrastruktur
- Feide og Kalmar2 har vist seg å være nyttig for brukerautentisering i CLARIN og INESS: en effektiv matrise av id-leverandører og tjenesteleverandører

Hva er nødvendig for enda mer fremskritt?



Watson: 2500 regnekjerner + 25 forskere

Skape ny kompetanse

CLARA: Marie Curie Initial Training Network

- rekruttering av 19 forskere
- organisering av 10 forskerkurs
- Samarbeid med lokale forskerskoler

Referanser

http://www.clarin.eu/

http://clara.uib.no/