Distributed Language Data Resources

Koenraad De Smedt

University of Bergen / CLARIN / DASISH

GN3 Innovation Workshop, Copenhagen, Oct. 10, 2011

Enormous amounts of language data exist

- The indexed Web contains about 40 billion pages (Oct 9, 2011; worldwidewebsize.com)
- 130 million books on Google Books (2010; 4% of all books)
- 13.6 billion words in Google Books published in 2008 alone
- The National Library of Norway digitizes 3000 to 4000 books every month (800 GB per day) in addition to collecting digital books, pictures, video etc.

Some things you can do with large language resources

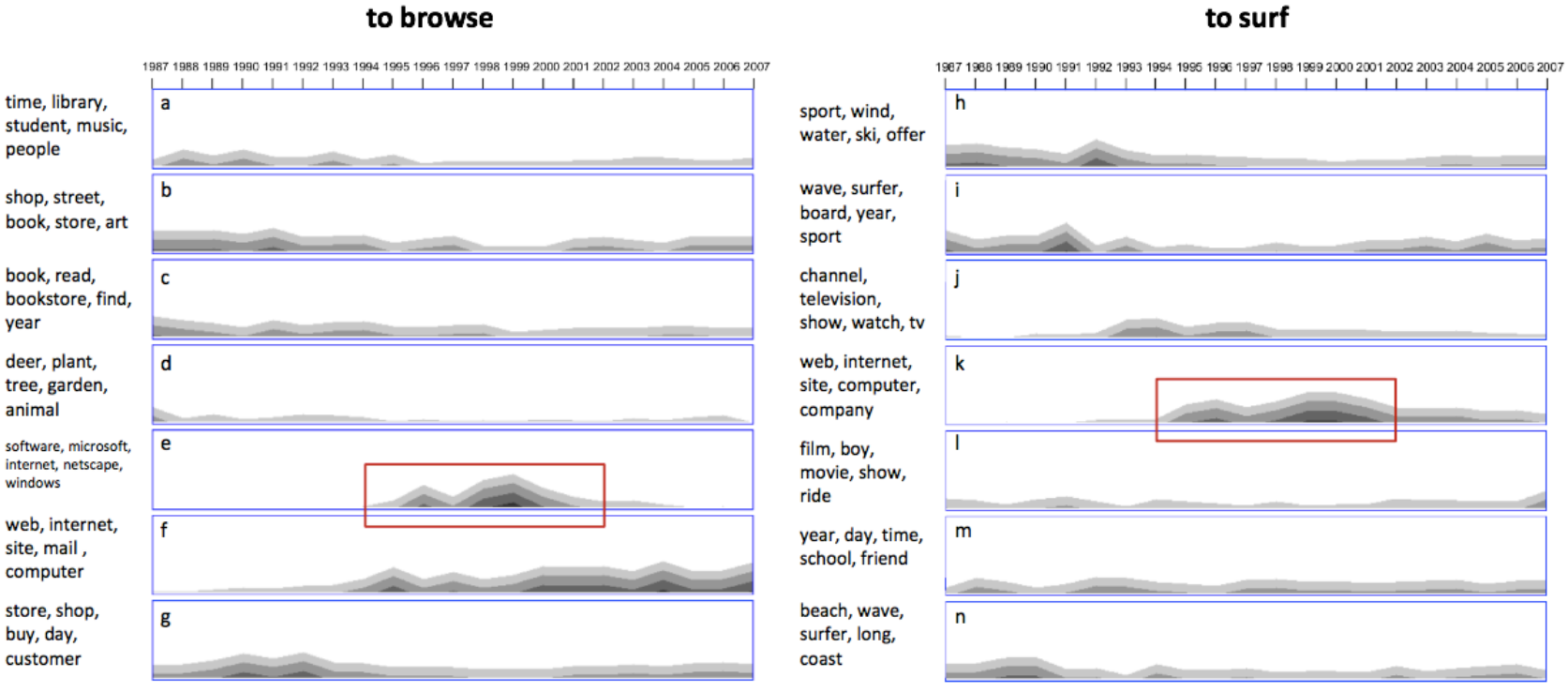
- Map words, phrases, quotes, ideas on time/geo
- Develop statistical models of grammar, translation, etc.
- Preserve and disseminate cultural heritage
- etc.

Aske 'ash' compounds, Spring 2010

In Norwegian web newspapers

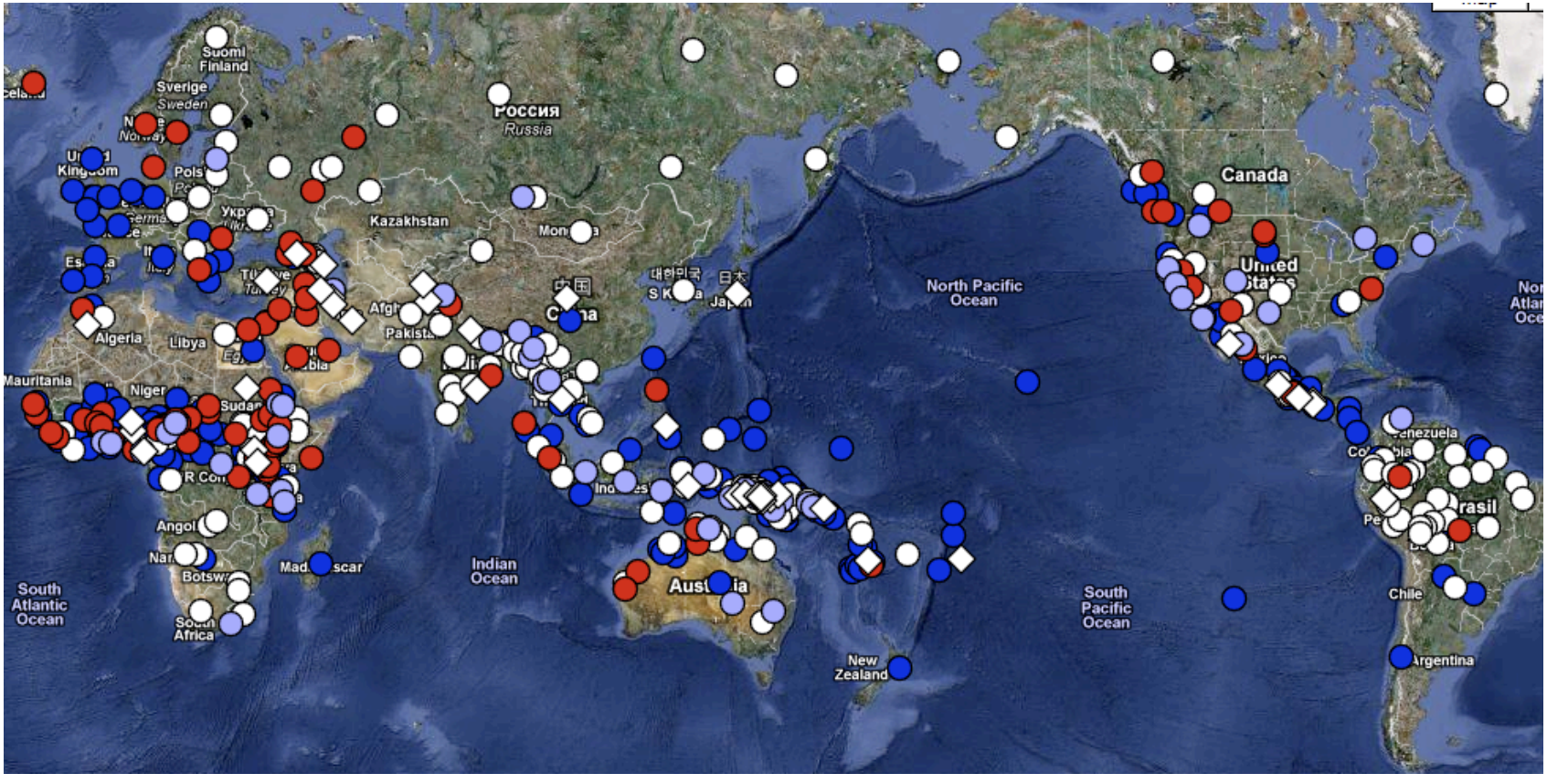
929 askeskyen
244 askesky
114 askefast
86 askeskyene
73 askeskyer
63 askefaste
32 askekaoset
30 askepartikler
26 askeproblemene
22 askekrisen
20 askekonsentrasjon
19 askefritt
18 askespredningen
18 askeskyens
17 askesituasjonen
16 asketrøbbel
16 askelaget
15 askeproblemer
15 askepartiklene
15 askefri
14 askelag
11 asketeppet
11 askestøv
10 asketrøbbelet
10 askeskya
9 askestansen
9 askekonsentrasjoner
8 askespredning
8 askenedfall
7 askeutslippet
7 askeprognosene
7 askeproduksjonen
7 askekaos
6 askeutsatte
6 asketap
6 askestøvet
6 askestrålen
6 askestoppen
6 askestans
6 askenedfallet
6 askebølge
6 aske-krisepakke
5 askevarsel
5 askestrømmen
5 askeregnet
5 askerammet
5 askeområdet
5 askegrå
5 askefylt
4 askevarslene
4 askeutslipp
4 askestengte
4 askeregn
4 askeprognose
4 askehumor
4 askehjelp
4 askefaren
4 aske-porteføljе
3 askevarselet
3 askeutslippene
3 asketomt
3 askesøylen
3 askesøyle
3 askestrandet
3 askestengt
3 askerammede
3 askenivået
3 askekartet
3 askehelgen
3 askefare
3 aske-strandede

Looking at changing meanings of words



Rohrdantz, Christian, Annette Hautli, Thomas Mayer, Miriam Butt, Frans Plank and Daniel A. Keim. 2011. Towards Tracking Semantic Change by Visual Analytics. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Short Papers), 305–310, 2011.

Marking of definiteness in world languages



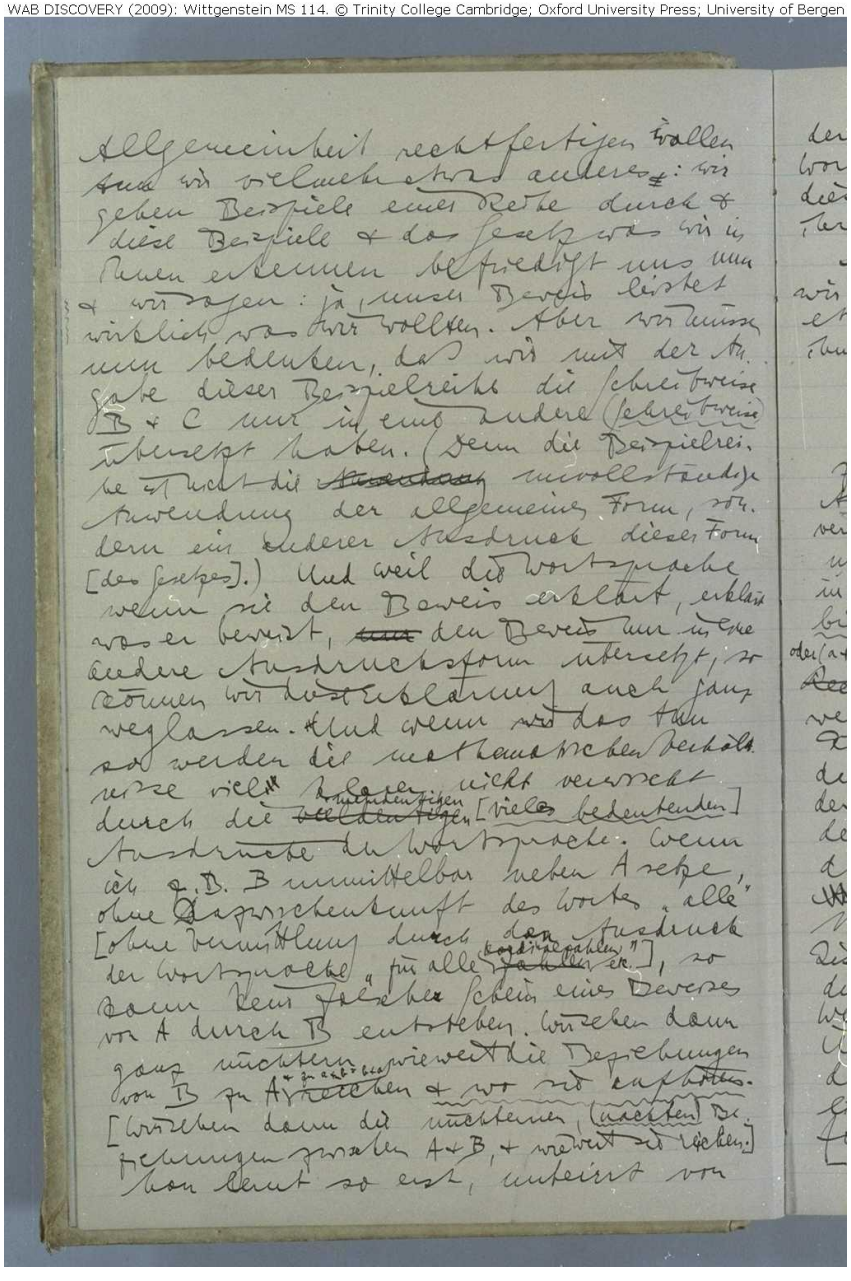
WALS: World Atlas of Language Structures

Curation of language resources

- Data integrity
- Annotation
- Rich metadata
- Access rights
- Advanced search
- Interoperability
- Permanence

Many valuable language data resources are scattered and ‘hidden’ to the research community

Wittgenstein archive: 1 of 40000 pages



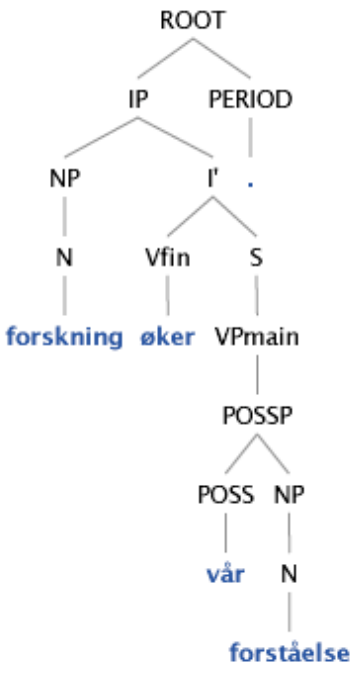
Allgemeinheit rechtfertigen **w**ollen
tun wir vielmehr etwas anderes ; <:> wir
gehen Beispiele einer Reihe durch &
diese Beispiele & das Gesetz was wir in
ihnen erkennen befriedigt uns nun
& wir sagen: ja, unser Beweis leistet
wirklich was wir wollten. Aber wir müssen
nun bedenken, daß wir mit der An-
gabe dieser Beispielreihe die Schreibweise
B & C nur in eine andere <(> Schreibweise <)>
übersetzt haben. (Denn die Beispielrei-
he ist nicht die ~~Anwendung~~ unvollständige
Anwendung der allgemeinen Form, son-
dern ein anderer Ausdruck dieser Form
[des Gesetzes] .) Und weil die Wortsprache
wenn sie den Beweis erklärt, erklärt
was er beweist, ~~nur~~ den Beweis nur in eine
andere Ausdrucksform übersetzt, so
können wir diese Erklärung auch ganz
weglassen. **U**nd wenn wir das tun
so werden die mathematischen Verhält-
nisse viel ~~<...>~~ klarer, nicht verwischt
durch die ~~vieldeutigen~~ mehrdeutigen [vie**les** bedeutenden]
Ausdrücke der Wortsprache. Wenn
ich z.B. B unmittelbar neben A
setze, ohne [~~d~~|**D**]azwischenkunft des Wortes „alle"
[ohne Vermittlung dur**c**h d[~~as~~|**en**] Ausdruck
der Wortsprache „für alle ~~Zahlen~~ ˇKardinalzahlen < etc. >"] , so
kann kein falsch**e**r Schein eines Beweises
von A durch B entstehen. Wir sehen dann
ganz nüchtern wie weit die Beziehungen
von B zu A ˇ& zu a + b = b + a reichen & wo sie aufhören.
[Wir sehen dann die nüchternen, <(> nackten <)> Be-
ziehungen zwischen A & B, & wie weit sie re<i>chen.]
Man lernt so erst, unbeirrt von

Treebanks: annotation much larger than source material

Parse sentence | Reload and parse | ☐ Packed representation | ☐ Disable Optimality marks | ☑ Suppress CHECK | ☐ Suppress complex categories | ☐ PREDs only | ☐ Show discriminant weights | ☐ Include non-top F-structures | Show ☐ discriminants ☑ c-structure ☑ f-structure ☐ MRS

1+1 solutions, 0.09 CPU seconds, 233 subtrees unified; rank: 0.0 (Only optimal solutions are shown.)

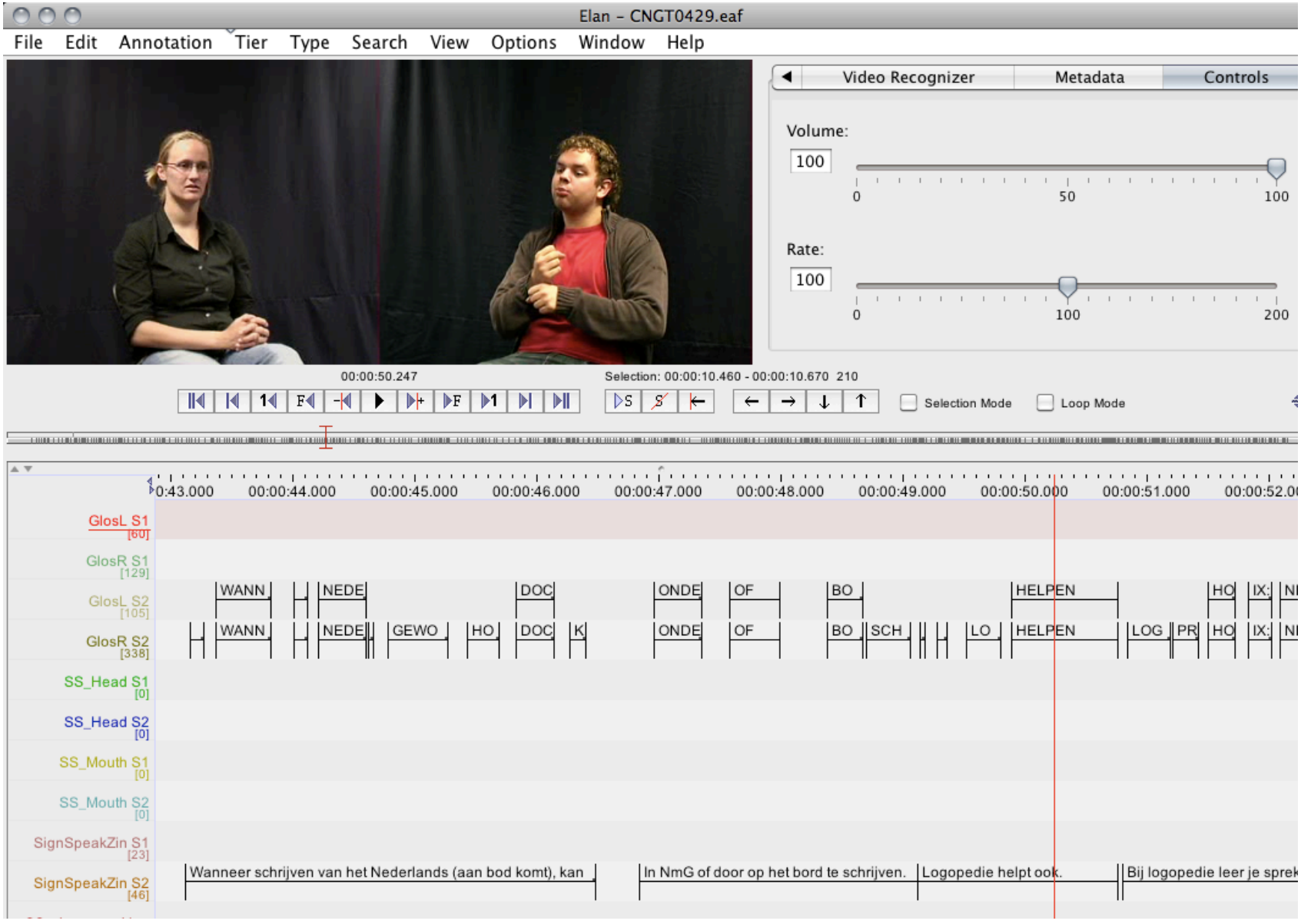
C-structure



F-structure

```
PRED      'øke<[12:forskning], [13:forståelse]>NULL'
TNS-ASP   15 | TENSE pres, MOOD indicative |
TOPIC     PRED   'forskning'
          NTYPE  | NSEM 21 | COMMON mass |
                 19 | NSYN common |
          GEND   18 | NEUT -, MASC +, FEM - |
          12 | PERS 3, NUM sg, CASE nom |
OBJ       PRED   'forståelse'
          SPEC   | POSS | PRED 'pro'
                 37   40 | POSS-TYPE 1pers,
                         | NUM pl |
          NTYPE  | NSEM 39 | COMMON count |
                 36 | NSYN common |
          GEND   35 | NEUT -, MASC +, FEM - |
          13 | REF +, PERS 3, NUM sg, DEF +, CASE obl |
SUBJ      [12]
0 | VTYPE main, VFORM fin, STMT-TYPE decl |
```

Annotated video: sign language corpora



Diverse user communities working with language data

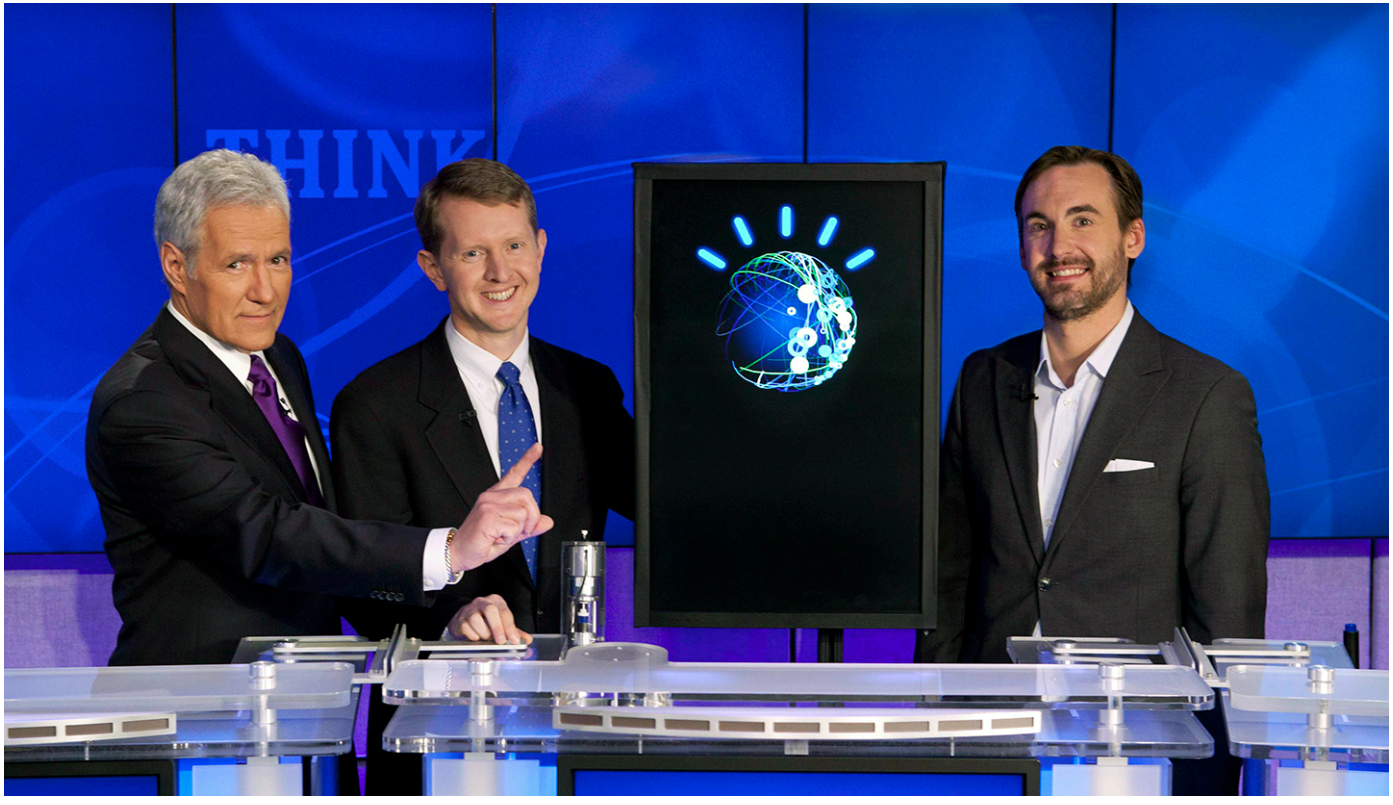
Linguistics and computational linguistics, literature, anthropology, history, philosophy and logic, computer science and artificial intelligence, psychology, neurology, education and pedagogy, …

Variety of data: source texts, translated/edited texts, syntactically and semantically annotated texts, speech and video recordings, transcriptions, concordances, parallel corpora, dictionaries, word nets, termbanks, grammars, eye tracking data, dialect maps, etc.

Many kinds of data $\times$ many different perspectives

Who's managing the data?

A special data integration for a particular purpose



IBM DeepQA (Watson) combines knowledge from many language resources with a UIMA-based system, developed in 4 years by 25 people on a HPC cluster with 2500 compute cores

Language data resources are scattered

- Different formats, standards, licenses, access mechanisms
- Inconsistencies due to lack of versioning
- Largely unstructured, not easily combinable or comparable
- Insufficiently indexed, insufficient metadata and documentation
- Unmanaged, invisible, not curated

More problems handling language data resources

- Primary data are often unmanaged, messy, noisy, unstructured, inconsistent, ambiguous
- Annotation needed (open-ended, many-layered, task dependent, not always reusable)
- Many obsolete formats and decaying information bearers
- Good analysis tools are missing, unreliable or not easily adaptable
- Few language data collections have a permanent source of funding
- Many language data are protected by copyright and/or encumbered by privacy considerations

Some digital language research needs

- Cooperation between stakeholders to share high quality data and make them visible and accessible
- Mashups combining materials and annotations from different resources/locations
- Workflows with tools from different sources, tailoring to data formats and user needs
- Partly static data, partly changing (reparsing, new annotations etc.)
- Respect rights, licences and privacy (can be quite complex)
- Provenance information, persistency (also for mashups)
- High need for specific visualizations and text representations
- Preserve and curate data which is often irreplaceable

eHumanities needs

- Humanities lack tradition for technical support which natural sciences have
- Currently a low number of users, but high potential
- Moderate but growing CPU cycle needs, growing storage and service needs
- High need for simplifying sharing of distributed data through user and licence management
- Training of the large potential user community
- Collaboration tools to go beyond one-way knowledge access
- Data sharing as publication

CLARIN

Common Language Resources and Technologies Infrastructure

“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.”

- Pan-European ESFRI project, 24 countries (preparatory phase 2008–2011)
- ERIC legal entity applied for; national commitments under way

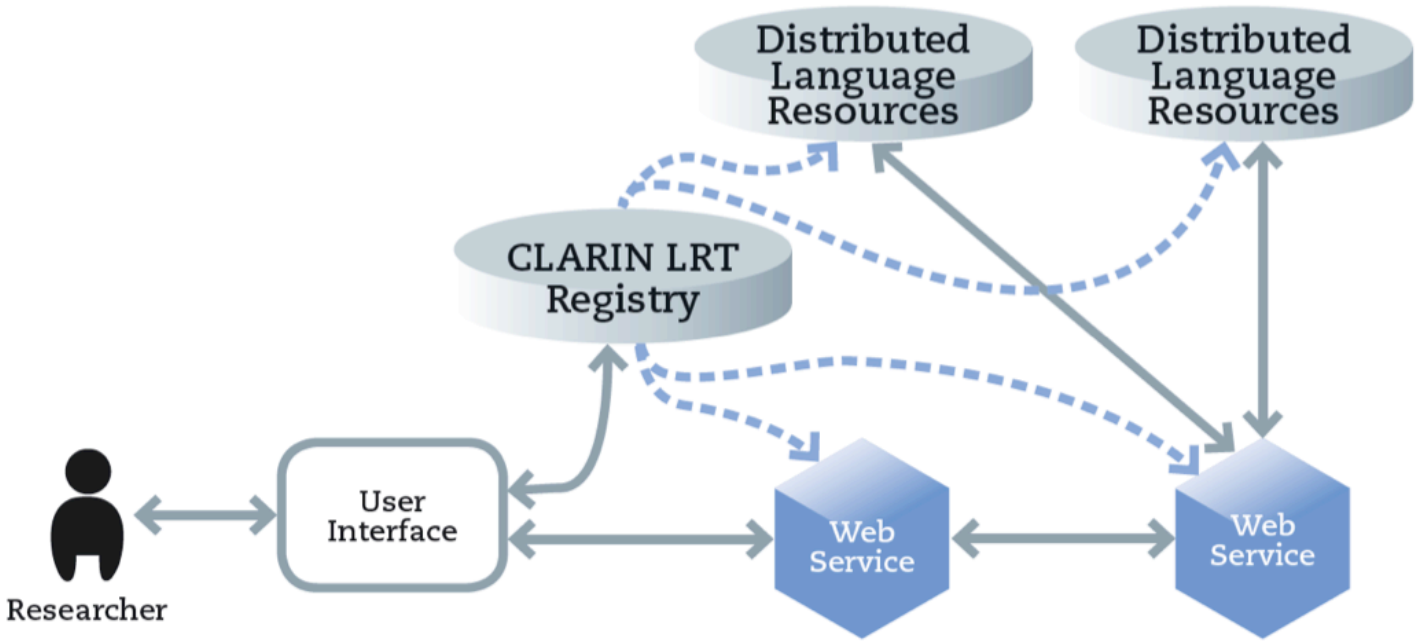
Usage scenario

- A researcher authenticates at her own organization and creates a “virtual” collection of resources from different repositories
- She does this on the basis of browsing a catalogue, searching through metadata, or searching in resource content
- To be granted access to this distributed dataset she signs the appropriate licenses
- She is then able to use a workflow specification tool and process this virtual collection using LT tools in the form of reliable distributed web services which he is authorized to use.
- Intermediate results are stored in a user specific workspace
- After evaluation, the resulting data (including metadata) can be added to a repository and the “virtual” collection specification can be stored for future reference using PIDs.

eInfrastructure components

- Trusted repositories for data and services e.g. CLARIN centers
- Metadata catalog for browsing and searching
- Virtual collection registries to store user specified collections and share them
- AAI infrastructure for technical, organizational, legal issues
- Distributed workspaces
- Persistent identification of resources to make references last

Web services



Joint metadata domain

- Currently a fragmented metadata landscape: DC, OLAC, IMDI, TEI Header, etc.
- Wrong terminology, too many or too few descriptors, limited interoperability, ...

CLARIN chose to develop CMDI (Component MetaData Infrastructure) and adhere to TC37/ISOcat

- Based on explicit syntax and semantics
- Not so much a single new metadata set or schema but rather a metadata infrastructure supporting several schemas
- Allow researchers to create a new schema and choose and define their own metadata descriptors
- Support for existing sets as IMDI, DC/OLAC

Metadata lifecycle

Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry

Create metadata schema from selection of existing components. Allow creation of new components if they have references to ISOcat

Search Service

Semantic Mapping

Relation Registry

ISOcat Concept Registry

DCMI Concept Registry

other Concept Registry

CLARIN Component Registry

Joint Metadata Repository

Metadata harvesting by OAI protocol

Metadata Repository

Metadata Repository

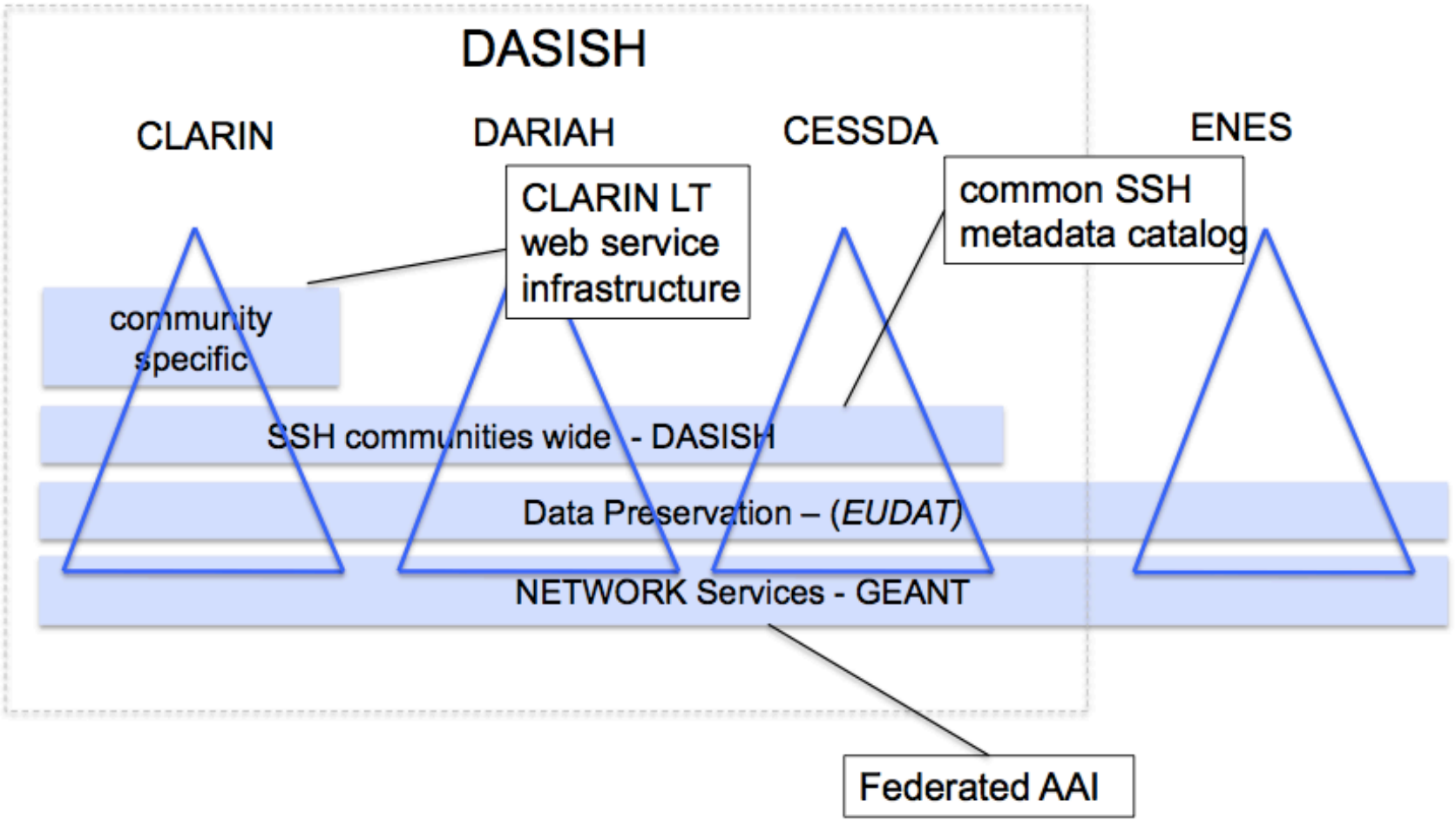
Metadata descriptions created

Metadata component profile was selected from metadata component registry

CLARIN layers

1. Coordination and governance layer (ERIC)
2. Infrastructure layer (long-term national responsibility)
3. Content creation layer (short-term projects by countries, institutions)

DASISH


DASISH
CLARIN
DARIAH
CESSDA
ENES
CLARIN LT
web service
infrastructure
common SSH
metadata catalog
community
specific
SSH communities wide - DASISH
Data Preservation – (EUDAT)
NETWORK Services - GEANT
Federated AAI

Some benefits of language curation and modeling

- Preservation of threatened language species
- Preservation of data on decaying bearers
- Management and much wider access to data

- Translation and cross-language information access
- Better interfaces, also inclusion of language impaired
- Deeper understanding of ideas, attitudes etc. in society/in history

eInfrastructure issues in distributed language data resources

How do community-specific and generic infrastructure layers work together?

How do we move from project-based infrastructure actions to very long term data archiving?

How do eLibraries, archives and national data services fit in the picture?

How will organizations such as EGI, PRACE and GEANT fit in offering grid, cloud, HPC and AAI services?

Acknowledgements

Thanks to Daan Broeder (MPI / CLARIN), NGT, WALS, the WAB and TREPIL projects for slide materials