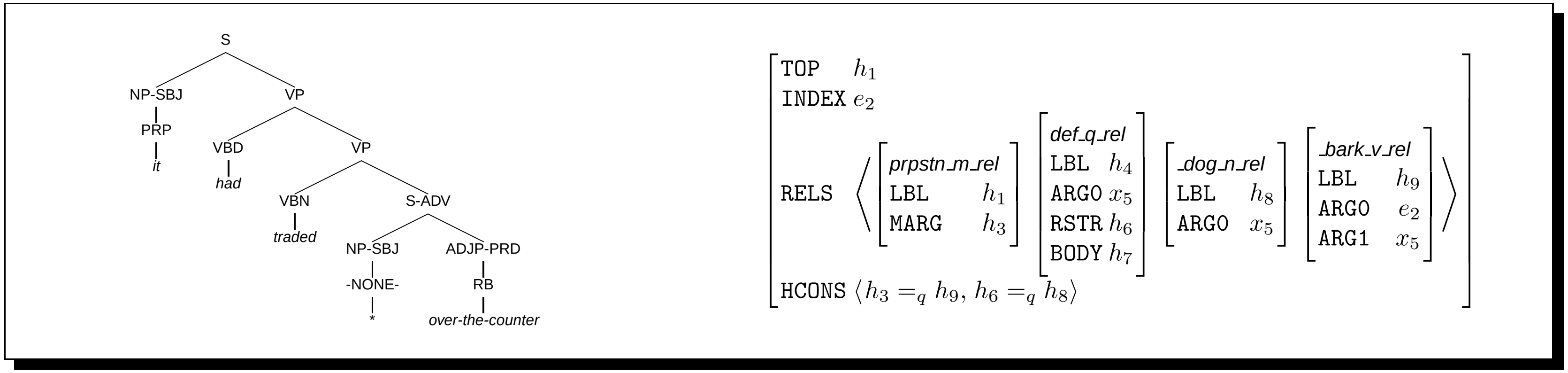


A Norwegian Language Grid

(A ‘Vision’ from the Semi-Outside)

Stephan Oepen

Universitetet i Oslo, Institutt for Informatikk

oe@ifi.uio.no

(CLARIN Norge — June 18, 2010)

D-SPIN: Language Resources & Technology On-Line

Pros and Cons of a Web-Based SOA:

⋮

– *Not applicable for huge amounts of data.*

⋮

The IFI Language Technology Group

Table of Contents

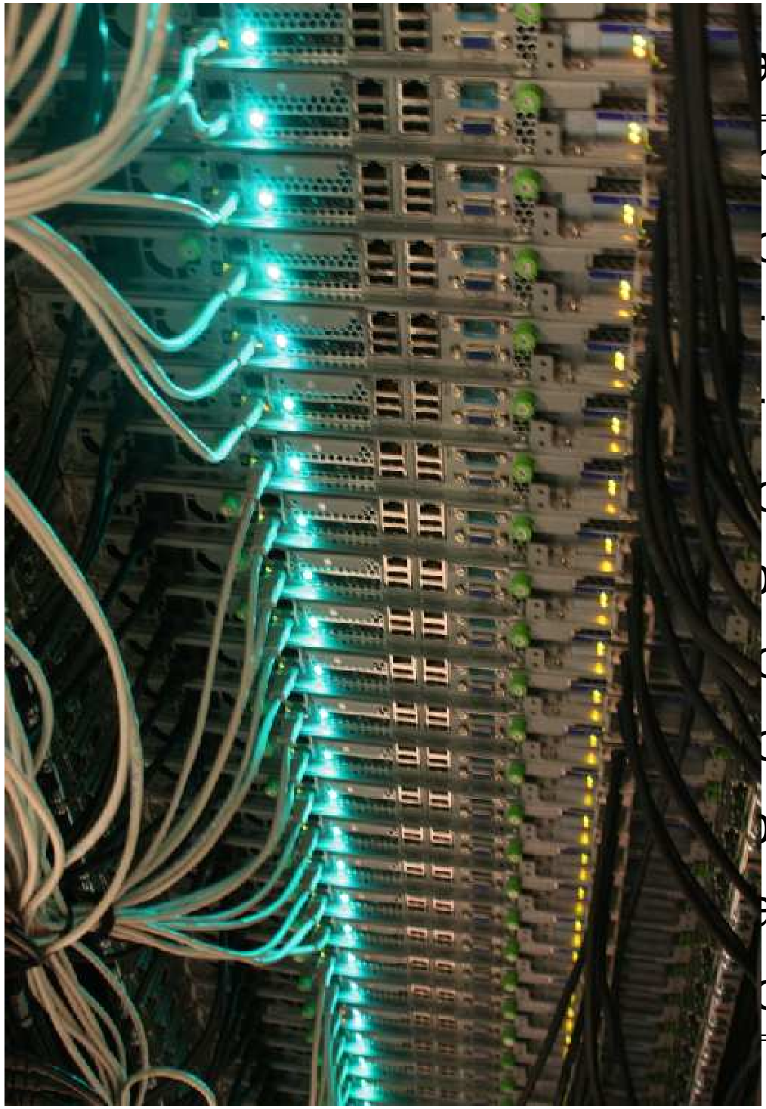
Gordana Ilić Holen	Doctoral Fellow	Coreference Resolution
Elisabeth Lien	Doctoral Fellow	Textual Inference
Jan Tore Lønning	Professor	Computational Semantics
Stephan Oepen	Professor	Grammar-Based Processing
Woodley Packard	Doctoral Fellow	Joint Disambiguation
Erik Velldal	Post-Doctoral Fellow	Classification
Gisle Ytrestøl	Doctoral Fellow	Incremental Parsing
Aleksander Øhrn	Adjunct Professor	Information Retrieval
Lilja Øvrelid	Post-Doctoral Fellow	Data-Driven NLP
NN	Associate Professor	Empirical Methods
NN	Doctoral Fellow	High-Quality Research

The IFI Language Technology Group

Table of Contents

Gordana Ilić Holen	Doctoral Fellow	Cor *Spring 2009* ution
Elisabeth Lien	Doctoral Fellow	Te *Fall 2009* ce
Jan Tore Lønning	Professor	Computational Semantics
Stephan Oepen	Professor	Grammar-Based Processing
Woodley Packard	Doctoral Fellow	Joi *Spring 2010* on
Erik Velldal	Post-Doctoral Fellow	*Fall 2009*
Gisle Ytrestøl	Doctoral Fellow	Incremental Parsing
Aleksander Øhrn	Adjunct Professor	Inf *Spring 2010* val
Lilja Øvrelid	Post-Doctoral Fellow	Da *Fall 2010* LP
NN	Associate Professor	En *Fall 2010* ods
NN	Doctoral Fellow	High *Fall 2010* earch

The IFI Language Technology Group



able of Contents

oral Fellow	Coreference Resolution
oral Fellow	Textual Inference
ofessor	Computational Semantics
ofessor	Grammar-Based Processing
oral Fellow	Joint Disambiguation
ctoral Fellow	Classification
oral Fellow	Incremental Parsing
ct Professor	Information Retrieval
ctoral Fellow	Data-Driven NLP
ate Professor	Empirical Methods
oral Fellow	High-Quality Research

An Example: Syntacto-Semantic Analysis of Wikipedia

General Idea

- Enabling technology: Wikipedia as a corpus and a knowledge source;
- e.g. research in linguistics, lexical acquisition, ontology learning, etc.

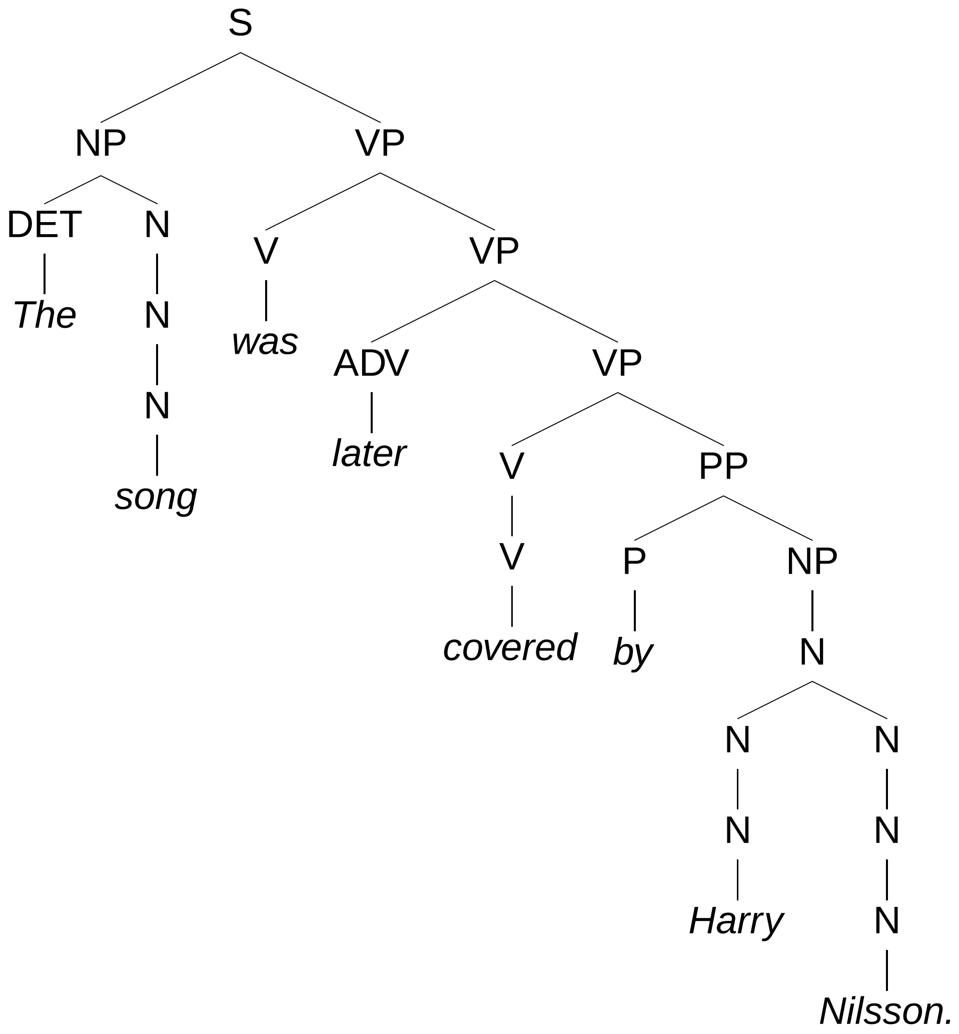
Approach & Technology

- Semi-automated ‘deep’ linguistic annotation, from pre-existing parser;
- gold-standard annotation of domain-specific subset: ~250,000 words.

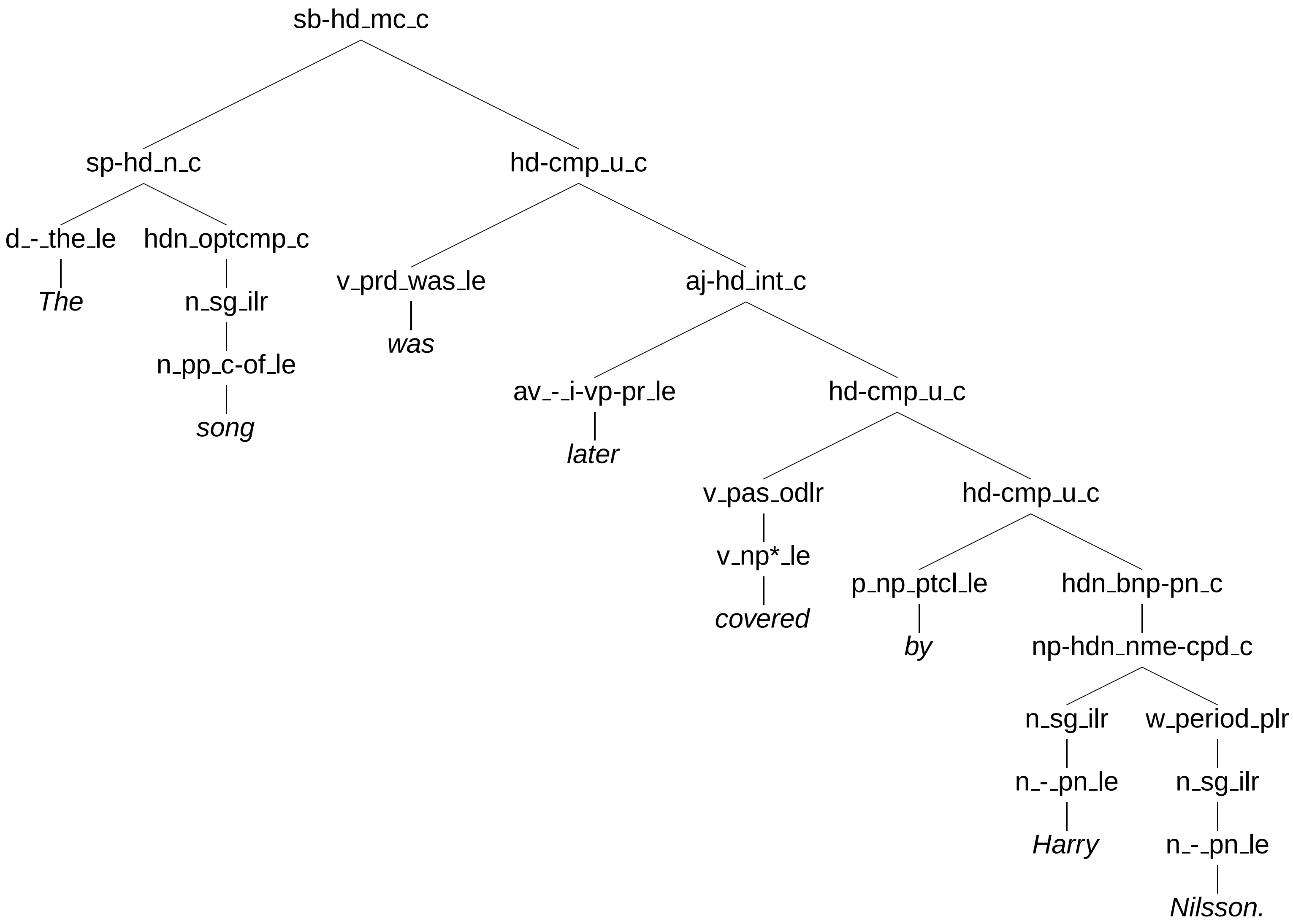
More Information (Download Site)

`http://www.delph-in.net/wikiwoods`

Syntactic Annotation: ‘Classic’ Constituent Tree



Syntactic Annotation: HPSG Derivation



Semantic Annotation: Predicate – Argument Structure

The song was later covered by Harry Nilsson.

$$
\langle\, h_1, \\
\left|\begin{array}{l}
h_3\text{:\_the\_q}(x_5, h_6, h_4), h_7\text{:\_song\_n\_of}(x_5\{\text{PERS } 3, \text{NUM } \textit{sg}\}, \_\_), \\
h_9\text{:\_cover\_v\_1}(e_2\{\text{SF } \textit{prop}, \text{TENSE } \textit{past}, \text{MOOD } \textit{ind}\}, x_{11}, x_5), \\
h_9\text{:\_later\_a\_1}(\_\_, e_2), \\
h_{16}\text{:compound\_name}(\_\_, x_{11}, x_{17}), \\
h_{19}\text{:proper\_q}(x_{17}, h_{20}, h_{21}), h_{22}\text{:named}(x_{17}\{\text{PERS } 3, \text{NUM } \textit{sg}\}, \textit{Harry}), \\
h_{13}\text{:proper\_q}(x_{11}, h_{14}, h_{15}), h_{16}\text{:named}(x_{11}\{\text{PERS } 3, \text{NUM } \textit{sg}\}, \textit{Nilsson})
\end{array}\right| \\
\{\, h_{20} =_q h_{22}, h_{14} =_q h_{16}, h_6 =_q h_7 \,\}\,\rangle
$$

Semantic Annotation: Predicate – Argument Structure

The song was later covered by Harry Nilsson.

$\langle\, h_1,$

h_3:_the_q(x_5, h_6, h_4), h_7:_song_n_of(x_5\{PERS 3, NUM *sg*\}, __),

h_9:_cover_v_1(e_2\{SF *prop*, TENSE *past*, MOOD *ind*\}, x_{11}, x_5),

h_9:_later_a_1(__, e_2),

h_{16}:compound_name(__, x_{11}, x_{17}),

h_{19}:proper_q(x_{17}, h_{20}, h_{21}), h_{22}:named(x_{17}\{PERS 3, NUM *sg*\}, *Harry*),

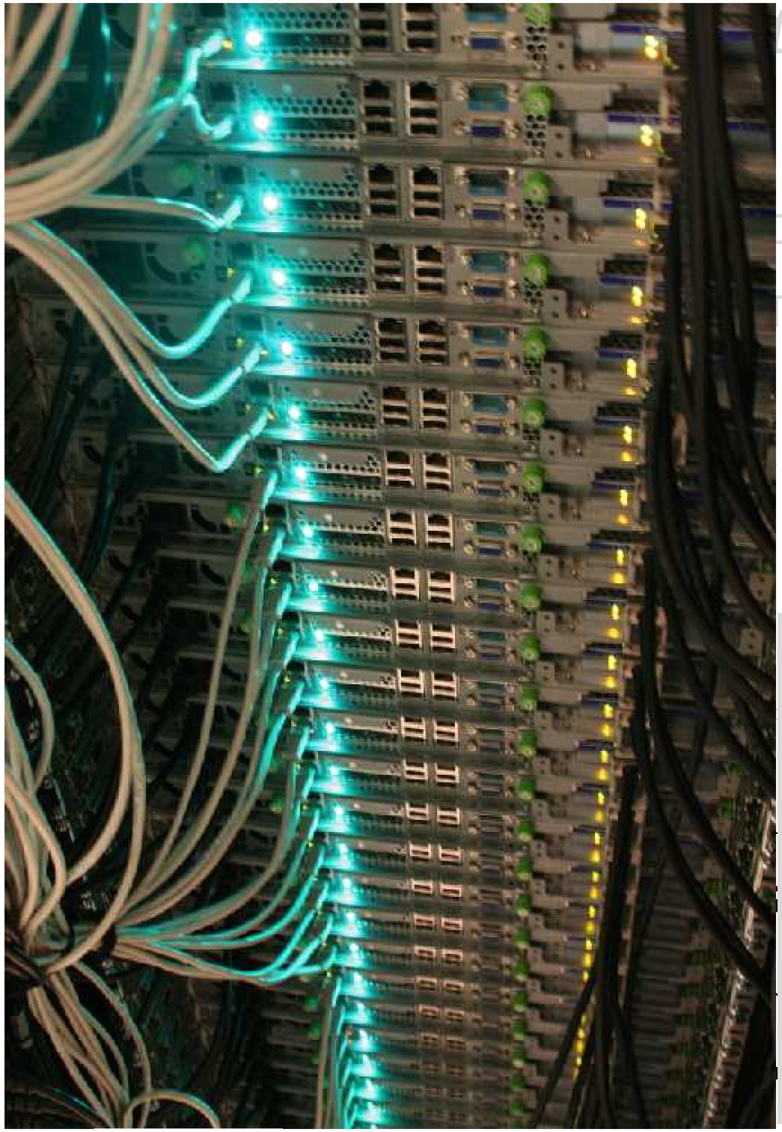
h_{13}:proper_q(x_{11}, h_{14}, h_{15}), h_{16}:named(x_{11}\{PERS 3, NUM *sg*\}, *Nilsson*)

$\{\, h_{20} =_q h_{22},\ h_{14} =_q h_{16},\ h_6 =_q h_7 \,\} \,\rangle$

→ 1.3 million content articles, 55 million utterances, ~900 million tokens;

→ ~85 % parsing coverage, ~83 % of analyses totally or nearly correct.

Semantic Annotation: Predicate – Argument Structure



ter covered by Harry Nilsson.

ong_n_of(x_5{PERS 3, NUM *sg*}, _),

TENSE *past*, MOOD *ind*}, x_{11}, x_5),

$_{11}$, x_{17}),

h_{22}:named(x_{17}{PERS 3, NUM *sg*}, *Harry*),

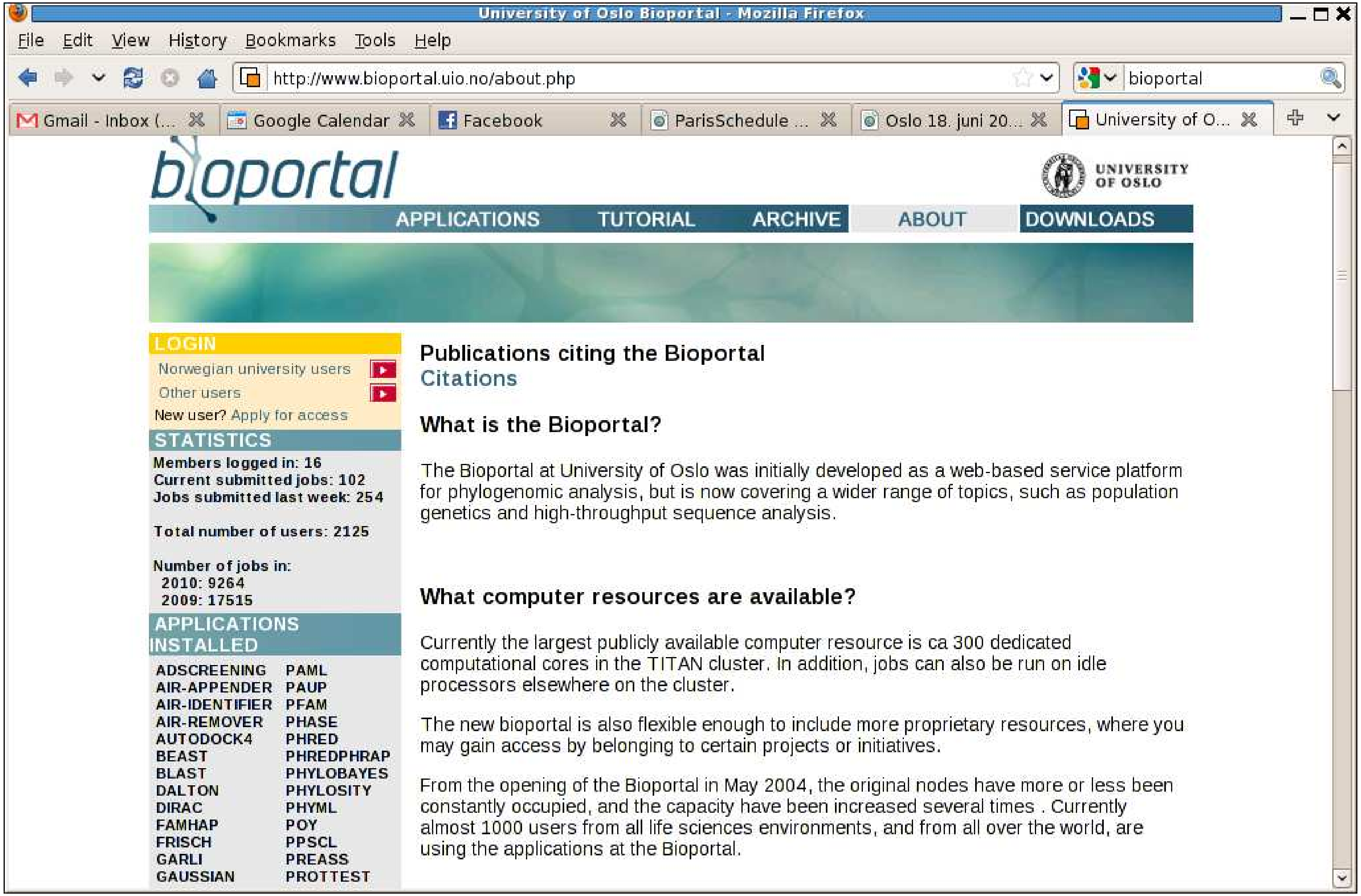
~120,000 cpu hours (six days);

~130 gigabytes compressed data;

→ subject extraction present in one of 15 utterances;

→ ~90 % in relative clauses.

A Candidate Role Model: BioPortal at UiO



LOGIN

Norwegian university users

Other users

New user? Apply for access

STATISTICS

Members logged in: 16
Current submitted jobs: 102
Jobs submitted last week: 254

Total number of users: 2125

Number of jobs in:
2010: 9264
2009: 17515

APPLICATIONS INSTALLED

ADSCREENING
AIR-APPENDER
AIR-IDENTIFIER
AIR-REMOVER
AUTODOCK4
BEAST
BLAST
DALTON
DIRAC
FAMHAP
FRISCH
GARLI
GAUSSIAN
PAML
PAUP
PFAM
PHASE
PHRED
PHREDPHRAP
PHYLOBAYES
PHYLOSITY
PHYML
POY
PPSCL
PREASS
PROTTEST

Publications citing the Bioportal
Citations

What is the Bioportal?

The Bioportal at University of Oslo was initially developed as a web-based service platform for phylogenomic analysis, but is now covering a wider range of topics, such as population genetics and high-throughput sequence analysis.

What computer resources are available?

Currently the largest publicly available computer resource is ca 300 dedicated computational cores in the TITAN cluster. In addition, jobs can also be run on idle processors elsewhere on the cluster.

The new bioportal is also flexible enough to include more proprietary resources, where you may gain access by belonging to certain projects or initiatives.

From the opening of the Bioportal in May 2004, the original nodes have more or less been constantly occupied, and the capacity have been increased several times . Currently almost 1000 users from all life sciences environments, and from all over the world, are using the applications at the Bioportal.

Imagine: Language Resources & Technology Portal

Motivation

- Reduce technology barriers: on-line demonstrators *and* processing;
- unified, Web-based point of entry; balance ease of use and flexibility.

Core Components

- **Data** *Språkbanken*, ELRA, LDC, and others; user-contributed data;
- **Tools** text extraction (PDF, {HT|W|X}ML, et al.), segmentation, morphology, tagging, chunking, parsing, search, concordancing, etc.

Scalability

- Built on top of national HPC infrastructure: NoTur, NorStore, NorGrid.

Imagine: Language Resources & Technology Portal

Motivation

- Reduce technology barriers: on-line demonstrators *and* processing;
- unified, Web-based point of entry; balance ease of use and flexibility.

Core Components

- **Data** *Språkbanken*, ELRA, LDC, and others; user-contributed data;
- **Tools** text extraction (PDF, {HT|W|X}ML, et al.), segmentation,

Preferably mostly through bottom-up, grass-roots process:
plurarilty of approaches: different frameworks and methods;
some convergence needed: exchange formats and interfaces;
starting points: UIMA, Language Grid, D-SPIN, and others.

More Concretely: Short-Term Initiatives

High-Performance LRT User Group

- UniNett Sigma (Σ) looking to establish discipline-specific user groups;
- group-internal functions: exchange experience, coordinate activities;
- interface function to Σ: give feedback on user experience and needs;
- at least one annual meeting → contact `oe@ifi.uio.no` if interested.

‘Deep’ Parsing Portal at UiO (`http://www.delph-in.net`)

- Existing international network on multi-lingual ‘deep’ parsing (HPSG);
- Fall 2010, seek NoTur support and cpu allocations to establish portal.

Credits

NoTur and NorStore (via UniNett Sigma);

The UiO Scientific Computation Group;

The Norwegian Taxpayers.