WikiWoods: Syntacto-Semantic Annotation for English Wikipedia

Dan Flickinger♣, Stephan Oepen♠, Gisle Ytrestøl♠

♣ Stanford University, Center for the Study of Language and Information
♠ University of Oslo, Department of Informatics

20 May 2010
LREC 2010, Malta

Annotation of the Wikipedia: Syntax and Semantics

Overview

- Extraction of relevant textual content
- Preprocessing and sentence segmentation
- Automatic parsing and disambiguation of full corpus
- Annotation export in several formats
- Manual annotation of NLP subcorpus



Motivation: Why Annotate the Wikipedia?

- Large on-line corpus of high-quality text
- Interesting and relevant content
- Mix of native and non-native authors
- Common annotation target: comparison and combination



Existing DELPH-IN Resources Used in WikiWoods

Consortium for deep linguistic processing resources:
`www.delph-in.net`

- Efficient parser: PET
- Broad-coverage, high-quality English Resource Grammar
- Minimal Recursion Semantics (MRS)
- Redwoods treebank annotation tools
- Statistical tools and methods for disambiguation

Format of Annotations for Each Sentence

Syntactic Analysis (HPSG)

- Full derivation tree
- Labeled with identifiers of constructions, lexical entries
- Recipe for constructing complete typed feature structure

Semantics (MRS)

- Fully linked graph of all elementary predications (relations)
- Head-argument and head-modifier dependencies
- Details of entities, events/states (e.g. number, aspect)
- Underspecified scope constraints

Format of Annotations for Each Sentence

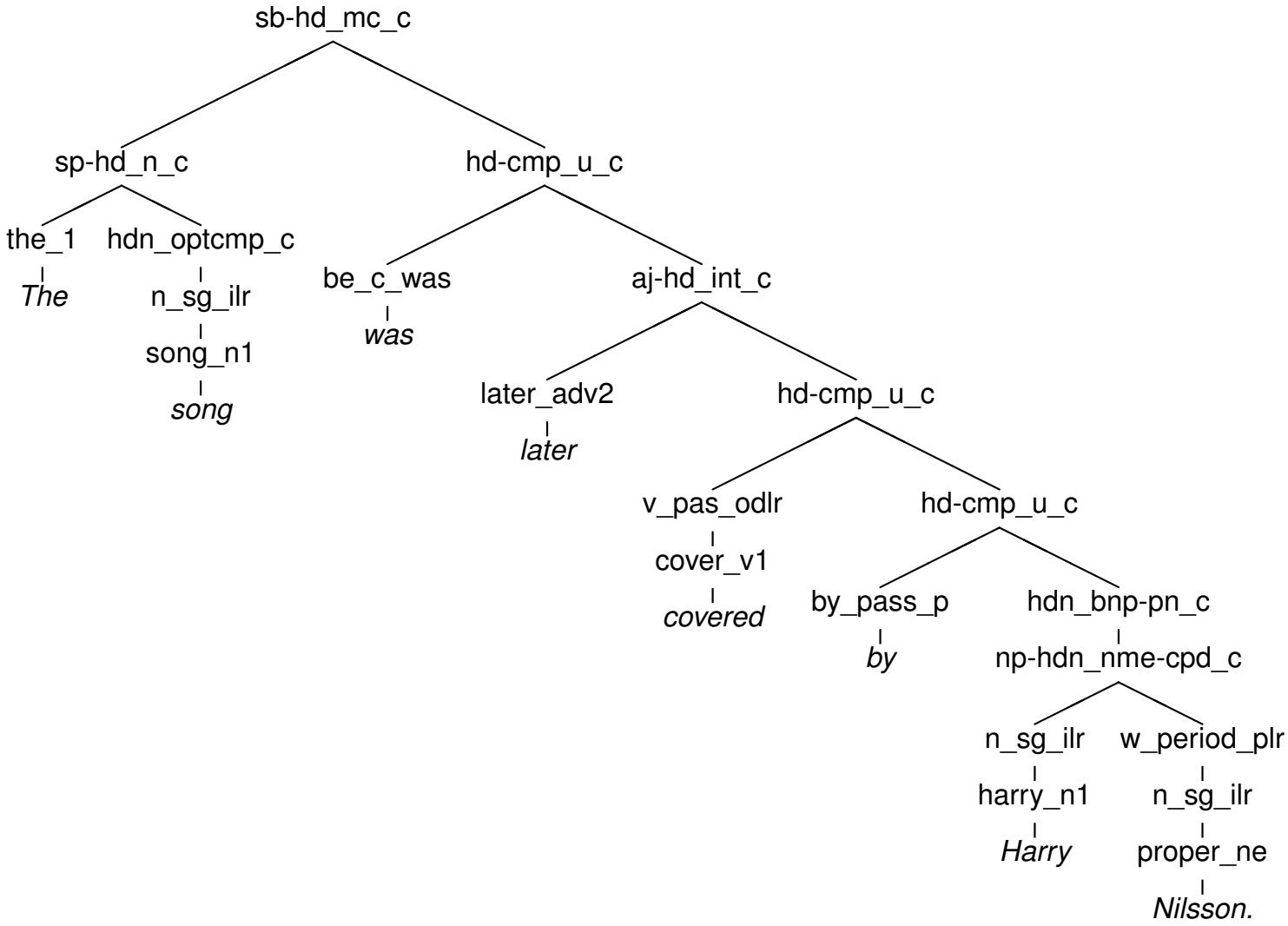
Syntactic Analysis (HPSG)

- Full derivation tree
- Labeled with identifiers of constructions, lexical entries
- Recipe for constructing complete typed feature structure

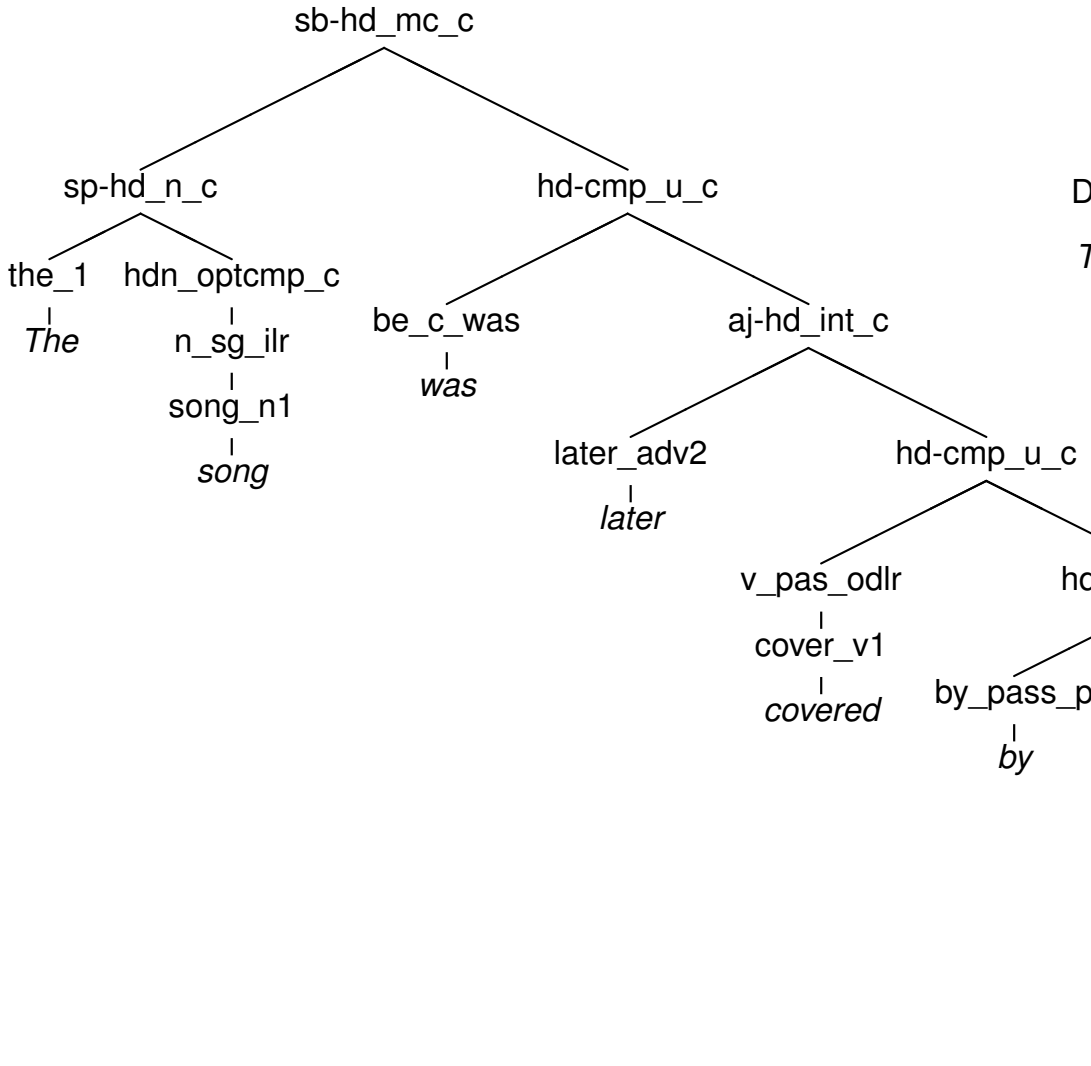
Semantics (MRS)

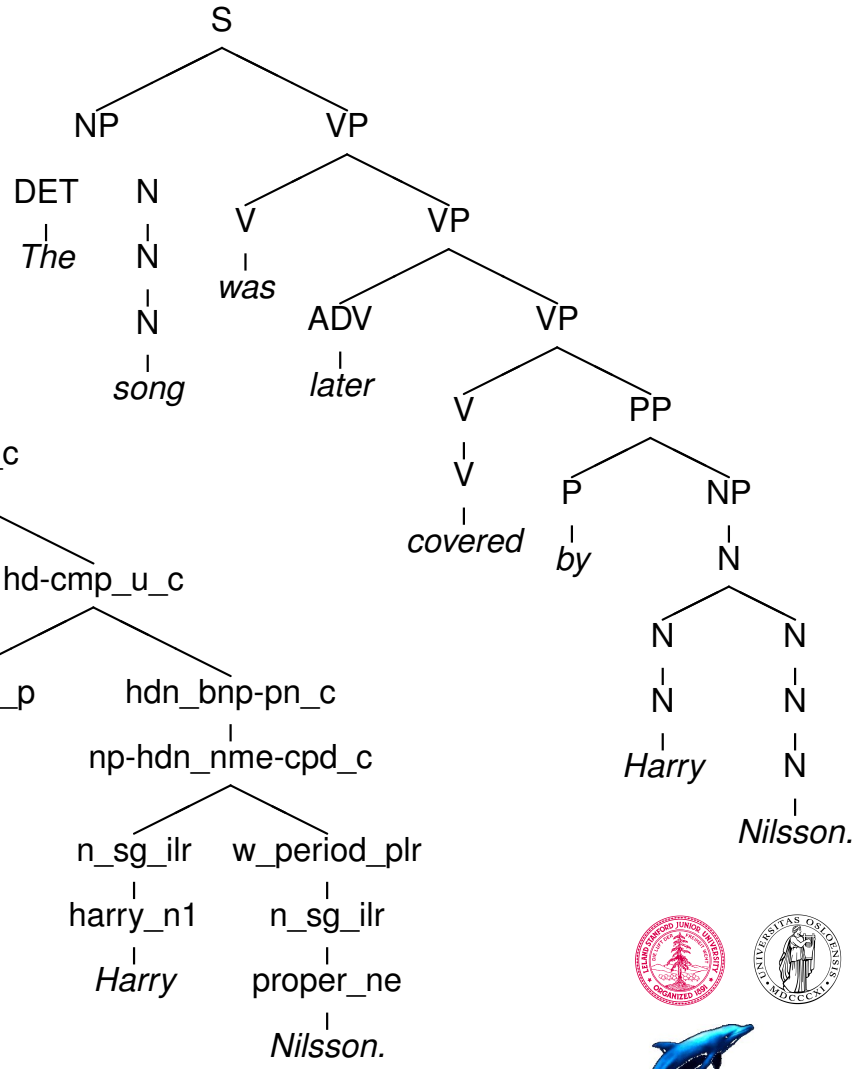
- Fully linked graph of all elementary predications (relations)
- Head-argument and head-modifier dependencies
- Details of entities, events/states (e.g. number, aspect)
- Underspecified scope constraints

Sample Syntactic Annotation



Sample Syntactic Annotation





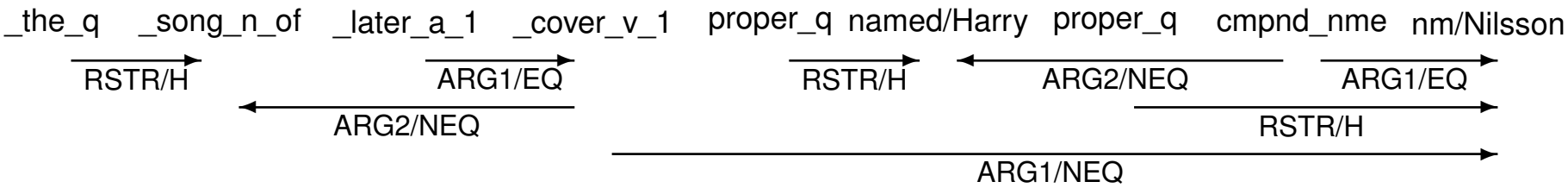
Sample Semantic Annotation

The song was later covered by Harry Nilsson.

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

Sample Semantics: Dependency MRS

The song was later covered by Harry Nilsson.



Method

Preprocessing

- Keep textual content, ‘linguistic’ markup (templates, fonts)
- Remove non-linguistic elements, including tabular content
- Use regular expression pattern-matching at textual level
- Produce output in plain text format, one sentence per line

Corpus organization

- Text files in segments of 100 consecutive articles
- Globally unique identifier for each utterance
- Several formats: *raw, text exchange, Redwoods treebank*

Method

Preprocessing

- Keep textual content, ‘linguistic’ markup (templates, fonts)
- Remove non-linguistic elements, including tabular content
- Use regular expression pattern-matching at textual level
- Produce output in plain text format, one sentence per line

Corpus organization

- Text files in segments of 100 consecutive articles
- Globally unique identifier for each utterance
- Several formats: *raw, text exchange, Redwoods treebank*

WeScience: Manually Annotated Subcorpus

Gold-standard disambiguation

- Higher-quality annotations
- Basis for estimates of expected error rates
- Training data for statistical parse disambiguation

Subcorpus size

- 100 articles on Natural Language Processing
- 16 segments, of which 3 are held out for testing
- 11,500 utterances in 13 annotated segments
- 10,100 parsed, 9,200 manually validated/disambiguated (80%)

WeScience: Manually Annotated Subcorpus

Gold-standard disambiguation

- Higher-quality annotations
- Basis for estimates of expected error rates
- Training data for statistical parse disambiguation

Subcorpus size

- 100 articles on Natural Language Processing
- 16 segments, of which 3 are held out for testing
- 11,500 utterances in 13 annotated segments
- 10,100 parsed, 9,200 manually validated/disambiguated (80%)

Scaling Up: The Complete Wikipedia

Corpus overview

- July 2008 snapshot
- Filtering out of very short articles, redirects, etc.
- 1.3 million articles, 55 M utterances, 900 million words

Automatic annotation

- Parse each utterance using PET and ERG with preprocessor/tagger
- Record most likely analysis (WeScience-trained model)
- Average ‘raw’ parse coverage at about 85%

Scaling Up: The Complete Wikipedia

Corpus overview

- July 2008 snapshot
- Filtering out of very short articles, redirects, etc.
- 1.3 million articles, 55 M utterances, 900 million words

Automatic annotation

- Parse each utterance using PET and ERG with preprocessor/tagger
- Record most likely analysis (WeScience-trained model)
- Average ‘raw’ parse coverage at about 85%

Sample Evaluation of Annotation Quality

- Random 1000 utterances from 500,000-utterance set
- Coarse-grained manual evaluation:
  correct: No errors in syntax or semantics
  nearly correct: One or two errors
  incorrect
- Roughly 82% receive correct or nearly correct analyses

Sample Evaluation of Annotation Quality

Item Length	Incorrect Parse	Nearly Correct	Correct Parse	Total Items
1 – 4	3	10	250	265
5 – 14	44	49	237	333
15 – 24	50	71	123	248
$\geq$ 25	50	51	47	154
Totals	**147**	**181**	**657**	**1000**

Outlook

- Full annotated corpus will be available this summer: `http://www.delph-in.net/wikiwoods/`
- Expect some 47 million annotated utterances
- Will adapt robust parsing methods to fill 15% gap in coverage
- Will continue validation and correction
- Expect use for e.g. information extraction, lexical semantics, ontology learning