

Motivation

- Internationalization: "Knowledge Society";
- Few speakers of Scandinavian languages;
- Keep Norwegian as an everyday language;
- Translators are costly, slow, and scarce;
- Machine translation is far from perfect, but + continuously improving through research, + cost-efficient, immediate, and abundant, + a useful tool for professional translators, + democratic: peer-to-peer communication, + typical language technology application.
- Practical testbed for foundational research.

Domain and Corpus

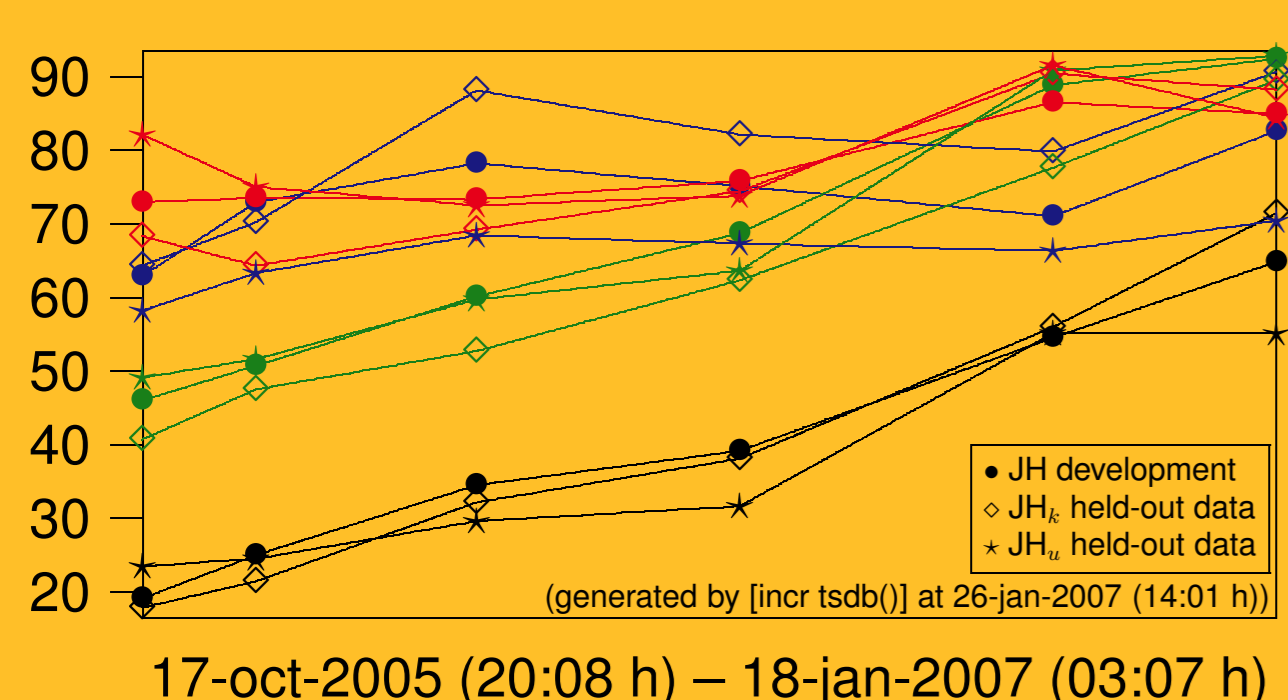
- Guidebooks to Norwegian backcountry, e.g. *På tur i Jotunheimen* (Per Roger Lauritzen);
- Target corpus of roughly 50,000 words;
- Two to three reference translations;
- Ten per cent held out for final evaluation.

Facts and Figures

- ~14,000 files; six programming languages;
- Coverage of ~6,000 Norwegian lexemes;
- Sponsored eight complete MSc theses;
- Six journal papers and six book chapters;
- 31 papers in peer-reviewed proceedings;
- 20 additional non-reviewed presentations;
- Co-authors spread out over four continents;
- Twelve visits from foreign researchers, ranging between three days and three months;
- Hosted four international events in Norway.

Development Process

- At least three full integrations each year;
- Daily regression evaluation and analysis of component and end-to-end performance.



Participants

University of Oslo

- Jan Tore Lønning;
- Stephan Oepen;
- Liv Ellingsen;
- Dan Flickinger;
- Kristin Hagen;
- Janne Bondi Johannessen;
- Lars Nygaard;
- Joel Priestley;
- Daniel Ridings;
- Erik Velldal.



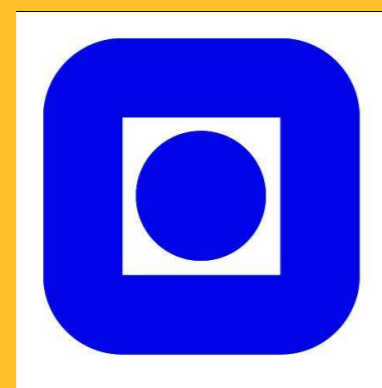
University of Bergen

- Helge Dyvik;
- Gunn Inger Lyse;
- Paul Meurer;
- Victoria Rosén;
- Sindre Sørensen.

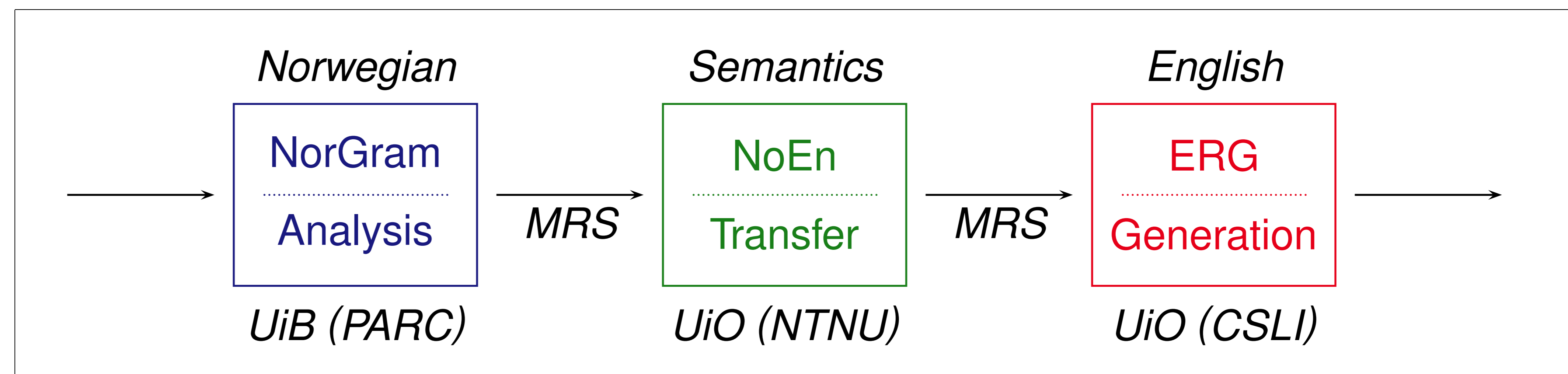


NTNU Trondheim

- Lars Hellan;
- Dorothee Beermann;
- Petter Haugereid;
- Torbjørn Nordgård.

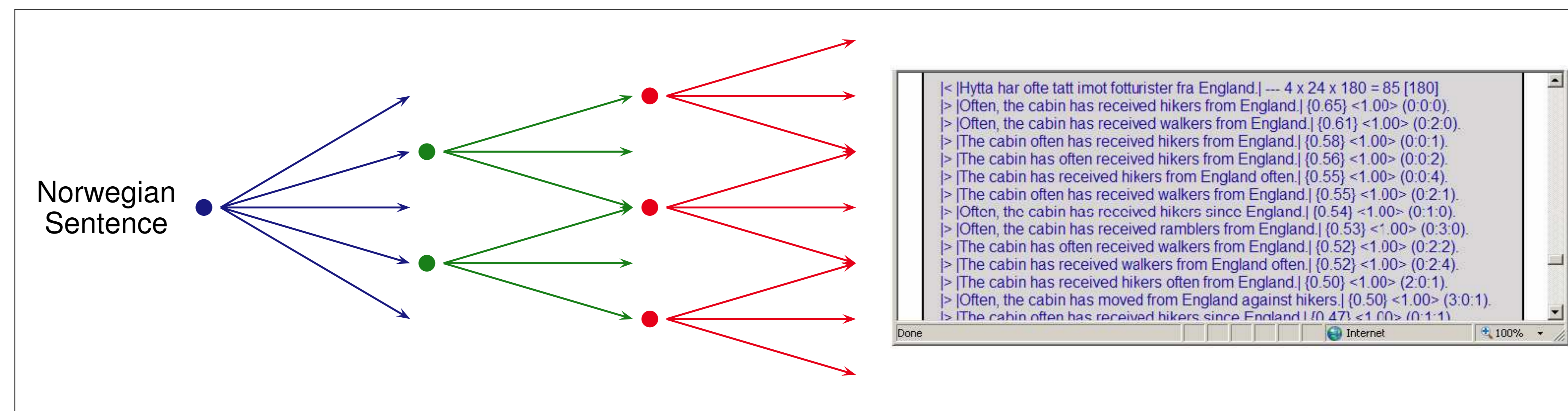


The Linguistic, Rule-Based Layer



Translation proceeds in three steps, each an independent module. First, a Norwegian sentence is grammatically analyzed, resulting in one or more logical-form representations of its meaning (in Minimal Recursion Semantics; MRS). Next, MRSs are transferred into English meaning representations. Finally, English sentences are generated from these representations. Each step builds on both (a) declarative linguistic specifications (e.g. the NorGram analysis grammar)—which are general-purpose resources, independent of the specific processing direction and application—and (b) specialized processing algorithms for analysis, transfer, and generation, respectively—putting the linguistic knowledge to work effectively.

The Stochastic Layer



At each step several different outputs may be produced. In analysis, for example, a string might be ambiguous, and in transfer there will often be alternative translations. At each step a ranking is applied to score competing hypotheses. These rankings are stochastic (i.e. frequentist) models trained from text corpora and treebanks. Per-component scores are calculated together. This yields a final score and (re-)ranking of alternative translations as shown with the output from the LOGON web demonstrator.

Analysis

The Norwegian grammar, NorGram, is developed within the framework of LFG at the University of Bergen. LOGON has extended the coverage (and efficiency) of the grammar considerably, and for the hiking domain the grammar now parses more than eighty per cent of all inputs.

As part of the project the grammar has been equipped with a novel semantic component producing MRSs. Also, the analysis component was extended in its morphology, now re-using existing tools for Norwegian (based on earlier collaboration between all three universities)—including compound analysis, PoS tagging, and NE recognition.

Transfer

A rewriting formalism for transfer of MRS meaning representations has been designed and implemented from scratch. The transfer formalism is fully typed and makes use of MRS unification. Transfer rules are applied non-deterministically, as there will often be multiple candidate translations for piece of source language MRS.

A core set of some 7,000 transfer rules are hand-coded, and there is at least one rule for each input lexeme. For simplex open-class lexemes (basic nouns and adjectives), an additional 10,000 transfer rules were acquired semi-automatically from a machine-readable dictionary.

Generation

For English, the English Resource Grammar (ERG) has been used. It has been under continuous development since 1994 by Dan Flickinger and colleagues at Stanford University. Dan joined the LOGON team at UiO for the second phase of the project, adapting the grammar to the tourism domain and helping harmonize meaning representations.

The generation component is built on top of the open-source DELPH-IN tools. As part of the project, in joint work with the University of Sussex, a 40-fold increase in generator efficiency was realized due to algorithmic refinements and a novel integration of stochastic rankings into the generator proper.

Doctoral Projects

Four doctoral stipendiaries were appointed in the project. Each of them has worked on a separate scientific sub-project—organized around the core demonstrator, though not mission-critical.



Liv Ellingsen researches non-categorical constraints on ordering of clausal elements, specifically different types of adverbials in English clauses. Based on corpus evidence, and proposing the inclusion of *soft constraints* into the generation grammar, her thesis aims for the most natural-sounding translations.



Gunn Inger Lyse has investigated a hybrid approach to Word Sense Disambiguation (WSD), combining classic statistical approaches and the Semantic Mirrors method. By inclusion of *lexico-semantic properties* in the WSD task, her project both improves the retrieval of relevant training data and overall WSD accuracy.



Petter Haugereid has revisited the division of labor between the lexicon and syntax components of a computational grammar of Norwegian. In a full-scale implementation of a novel *syntactic approach to linking*, his thesis accomplishes an underspecified lexical account of many types of argument structure variation.



Erik Velldal has developed rich stochastic models to rank alternate hypotheses in realization and transfer. Replacing the conventional approach to realization ranking with a *conditional structural model* and Maximum Entropy or SVM machine learning, ranker performance can be substantially improved.

Main Results

- Functional proof-of-concept demonstrator;
- New techniques for integration of linguistic and stochastic approaches → hybrid MT;
- Scalable MRS-based MT architecture, in use already in Japan, USA, and Germany;
- Methodological advances: profiling for MT;
- A comprehensive computational grammar of Norwegian, with domestic morphology;
- Novel transfer design and implementation;
- Greatly improved processing algorithms (received IJCNLP 2005 Best Paper Award).

Reusable Resources

- A high-precision analyzer for Norwegian;
- Morphology, PoS tagger, NE recognizer;
- General-purpose MRS rewrite engine;
- Bi-lingual, aligned corpus and treebanks;
- Cross-language computational semantics.