



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Statistical Techniques for Translating to Morphologically Rich Languages (Dagstuhl Seminar 14061)

**Citation for published version:**

Fraser, AM, Knight, K, Koehn, P, Schmid, H & Uszkoreit, H 2014, 'Statistical Techniques for Translating to Morphologically Rich Languages (Dagstuhl Seminar 14061)', *Dagstuhl Reports*, vol. 4, no. 2, pp. 1-16.  
<https://doi.org/10.4230/DagRep.4.2.1>

**Digital Object Identifier (DOI):**

[10.4230/DagRep.4.2.1](https://doi.org/10.4230/DagRep.4.2.1)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

Dagstuhl Reports

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Statistical Techniques for Translating to Morphologically Rich Languages

Edited by

Alexander Fraser<sup>1</sup>, Kevin Knight<sup>2</sup>, Philipp Koehn<sup>3</sup>, Helmut Schmid<sup>4</sup>, and Hans Uszkoreit<sup>5</sup>

1 LMU München, DE, [fraser@cis.uni-muenchen.de](mailto:fraser@cis.uni-muenchen.de)

2 University of Southern California, USA, [knight@isi.edu](mailto:knight@isi.edu)

3 University of Edinburgh, GB, [pkoehn@inf.ed.ac.uk](mailto:pkoehn@inf.ed.ac.uk)

4 LMU München, DE, [schmid@cis.uni-muenchen.de](mailto:schmid@cis.uni-muenchen.de)

5 Universität des Saarlandes, DE, [uszkoreit@coli.uni-sb.de](mailto:uszkoreit@coli.uni-sb.de)

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14061 “Statistical Techniques for Translating to Morphologically Rich Languages”. The seminar took place in February 2014. The purpose of the seminar was to allow disparate communities working on problems related to morphologically rich languages to meet to discuss an important research problem, translation to morphologically rich languages. While statistical techniques for machine translation have made significant progress in the last 20 years, results for translating to morphologically rich languages are still mixed versus previous generation rule-based systems, so this is a critical and timely topic. Current research in statistical techniques for translating to morphologically rich languages varies greatly in the amount of linguistic knowledge used and the form of this linguistic knowledge. This varies most strongly by target language, for instance the resources currently used for translating to Czech are very different from those used for translating to German. The seminar met a pressing need to discuss the issues involved in these translation tasks in a more broad venue than the ACL Workshops on Machine Translation, which are primarily attended by statistical machine translation researchers. The report describes the introductory material presented to the group, the organization of break-out discussion groups by topic, and the results of the seminar.

**Seminar** February 2–7, 2014 – <http://www.dagstuhl.de/14061>

**1998 ACM Subject Classification** I.2.7 Natural Language Processing – Machine Translation

**Keywords and phrases** Machine Translation, Statistical Machine Translation, Syntactic Parsing, Morphology, Machine Learning, Morphologically Rich Languages

**Digital Object Identifier** 10.4230/DagRep.4.2.1



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Statistical Techniques for Translating to Morphologically Rich Languages, *Dagstuhl Reports*, Vol. 4, Issue 2, pp. 1–16

Editors: Alexander Fraser, Kevin Knight, Philipp Koehn, Helmut Schmid, and Hans Uszkoreit



DAGSTUHL Dagstuhl Reports

REPORTS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary


*Alexander Fraser*

*Kevin Knight*

*Philipp Koehn*

*Helmut Schmid*

*Hans Uszkoreit*

**License**  Creative Commons BY 3.0 Unported license  
© Alexander Fraser, Kevin Knight, Philipp Koehn, Helmut Schmid, and Hans Uszkoreit

This report documents the program and the outcomes of Dagstuhl Seminar 14061 “Statistical Techniques for Translating to Morphologically Rich Languages”. The website of the seminar, which allows access to most of the materials created for and during the seminar, is <http://www.dagstuhl.de/14061>. The seminar on Statistical Techniques for Translating to Morphologically Rich Languages allowed disparate communities working on problems related to morphologically rich languages to meet to discuss an important research problem, translation to morphologically rich languages. While statistical techniques for machine translation have made significant progress in the last 20 years, results for translating to morphologically rich languages are still mixed versus previous generation rule-based systems, so this is a critical and timely topic. Current research in statistical techniques for translating to morphologically rich languages varies greatly in the amount of linguistic knowledge used and the form of this linguistic knowledge. This varies most strongly by target language, for instance the resources currently used for translating to Czech are very different from those used for translating to German. The seminar met a pressing need to discuss the issues involved in these translation tasks in a more broad venue than the ACL Workshops on Machine Translation, which are primarily attended by statistical machine translation researchers.

Important background for the discussion was the recent realization that more linguistically sophisticated methods are required to solve many of the problems of translating to morphologically rich languages. Therefore it was critically important that SMT<sup>1</sup> researchers be able to interact with experts in statistical parsing and morphology who work with morphologically rich languages to discuss what sort of representations of linguistic features are appropriate and which linguistic features can be accurately determined by state of the art disambiguation techniques. This was an important step in creating a new community crossing these research areas. Additionally, a few experts in structured prediction were invited. The discussions took advantage of their insight in how to jointly model some of these phenomena, rather than combining separate tools in ad-hoc pipelines as is currently done. The overall discussion was driven by the following questions:

- Which linguistic features (from syntax, morphology and other areas such as coreference resolution) need to be modeled in SMT?
- Which statistical models and tools should be used to annotate linguistic features on training data useful for SMT modeling?
- How can we integrate these features into existing SMT models?
- Which structured prediction techniques and types of features are appropriate for training the extended models and determining the best output translations?
- What data sets should be used to allow a common test bed for evaluation?

---

<sup>1</sup> SMT – Statistical Machine Translation

- How should evaluation be conducted, given the poor results of current automatic evaluation metrics on morphologically rich languages?

The Dagstuhl seminar on Statistical Techniques for Translating to Morphologically Rich Languages addressed these questions by allowing four different communities to meet together: statistical machine translation, statistical parsing, morphology and structured prediction.

**Outcome in brief.** The Dagstuhl seminar on Statistical Techniques for Translating to Morphologically Rich Languages was a great success. The discussions held will play an important role in allowing researchers to significantly advance the state-of-the-art. In particular, strong and weak points in current research approaches were identified and proposals to address the weak points were made. In addition, the seminar acted as a valuable venue for more junior researchers to spend more time talking with senior researchers than is possible in a conference setting. Finally, several new community building ideas were discussed, including a DFG proposal connecting all of the major sites for statistical machine translation research in Germany, see below.

**Invited Talks.** We begin the detailed discussion with a brief idea about the three invited keynote talks (as well as the introductory overview and motivational talk). All of these talks were very well received, with several seminar participants commenting that they learned a significant amount by being able to see a synthesis of the problems, current approaches and possible future approaches to translating to morphologically rich languages. The three keynote talks were:

- Philipp Koehn of the University of Edinburgh presented a general discussion of dealing with the phenomena of morphologically rich languages in translation.
- Kristina Toutanova of Microsoft Research presented a detailed overview of the state-of-the-art in statistical machine translation research related to morphologically rich languages in translation.
- Kevin Knight of the University of Southern California presented a vision of the future, where the field could go, in terms of both better modelling of morphologically rich languages, and the use of more language independent structure (at the semantic level) in translation.

After this, people interested in leading a discussion group held talks.

**Discussion Groups.** There were initially nine proposed topics for discussion groups (note that these are listed as topic-focused talks subsequently in the report):

- Nivre/Petrov: Parallel dependency treebanks and linguistic resources
- Tiedemann: The use of synthetic training data and pivot languages to overcome data sparseness
- Kirchhoff: Language modeling
- Dyer: Modeling morphemes vs. modeling words and smoothing with morphemes
- Habash: Arabic morphology and deep morphology representation for MT
- Williams/Koehn: Syntactic SMT for morphologically rich languages
- Knight: Semantics
- Webber: Discourse/aspects of semantics
- Bojar/Hajič: Generating morphology for SMT

Following this all participants emailed the organizers with their discussion group preferences. In the end, all but two participants were assigned to their first preference. We eliminated two groups (on synthetic training data and generating morphology), and their proposers joined other groups.

Following initial group presentations by some groups on Wednesday morning, three groups dissolved and several decided to continue. The three new groups that were proposed were:

- Virpiojia/Dyer: Unsupervised morphology for statistical machine translation
- Wu/Lavie: Evaluation of machine translation output
- Nivre/Knight: Universal Annotation and Abstract Meaning Representation

Highlights of what was accomplished by the discussion groups were:

- Dyer and Virpiojia and groups looked at morphologically aware translation models which use morphology to cover the long-tail without requiring morphological modelling of very frequent tokens, and looked at the state-of-the-art in unsupervised modeling.
- Kirchhoff and her group carried out a detailed survey of the state-of-the-art for language modeling of morphologically rich languages and documented this on the Wiki.
- Nivre and his two groups (one co-led with Petrov) defined a new proposed annotation standard for working on two levels (surface forms and lemmas, including multi-word-entities and decomposed compounds).
- Habash and his group carried out a literature review of attempts to deal with Arabic morphology in translation, discussing the strengths and weaknesses of the approaches, and identifying a new direction for future work.
- Williams, Koehn and group looked at the application of unification to modelling agreement in multiple languages.
- Knight and his two groups worked on general applications of semantically-aware processing to morphologically rich languages and on identifying areas where the Abstract Meaning Representation could be applied to this problem.
- Webber and group created a list of resources and research papers on applying discourse modeling to statistical machine translation and looked at machine translation output to find errors caused by broken discourse constraints.
- Wu, Lavie and group discussed and documented the different levels of linguistic analysis required for high quality automatic evaluation when the target language is morphologically rich.

See the individual abstracts for more information and further details.

**Other activities.** In addition to the formal work carried out in the talks and discussion groups, Dagstuhl offered an intimate environment strongly encouraging networking and discussion. The meal system of Dagstuhl, with random assignment of people to tables, is an excellent idea and was particularly useful for the more junior participants who did not know many of the senior researchers attending (several people mentioned informally that this was the best experience of this sort they have had). The informal evening activities centering around social gatherings and the music room were also very well attended and a variety of interesting discussions took place. The excursion to Trier was a welcome mid-week break and provided another networking opportunity, as well as being highly interesting for the vast majority of participants who had not previously visited a city with a similar historical background.

The seminar was unusual for Dagstuhl itself in that very few of the participants had participated in a Dagstuhl seminar previously. Due to the strongly positive reaction we anticipate that other research areas within Natural Language Processing will apply for Dagstuhl seminars.

We would like to take the opportunity here to thank Dagstuhl for the wonderful logistic support and for providing such a stimulating environment for our work.

**Communities represented in more detail.** The seminar was a success in terms of the strong participation of women and a good geographical distribution (although Asia could have been somewhat more strongly represented). Our only strong area of concern was that of the numerous participants from companies invited, only two attended (Kristina Toutanova of Microsoft Research and Slav Petrov of Google, who gave one of the keynotes and co-led a discussion group respectively). Nevertheless the networking opportunities were excellent and many participants informally told us that this was an excellent meeting which they expected to have a strong impact on their research.

One characteristic of the proposal which was successfully carried out was a meeting of four different communities: statistical machine translation, statistical parsing, morphology and structured prediction. In particular, we felt that the interaction between the statistical machine translation researchers and the researchers working on statistical parsing and morphology was highly productive and will likely lead to new techniques of analyzing morphologically rich languages which will be more useful in translation research than the current approaches. We believe that the Dagstuhl seminar has been unique in terms of providing the opportunity for these communities to meet together for five days and understand each others' perspective on research.

**Conclusion and Impact.** In conclusion, we believe the Dagstuhl seminar has met the goals we set out for it, in terms of providing a forum for discussion of the current problems with the state-of-the-art and allowing a focusing of research effort which was not previously present in the research community.

As we previously mentioned, in addition to the less quantifiable aspects in terms of networking and connections made, there were several prominent concrete outcomes of the Dagstuhl seminar. The new annotation standard suggested by the two Universal Annotation groups led by Nivre, Petrov and Knight is one strong outcome which will change the basic tools that the statistical machine translation community will have available. The Kirchhoff group is working on a position paper that will help to refocus effort on language modeling for morphologically rich languages, which will have an impact not only on machine translation research but also research on speech recognition and other research areas.

Five of the six most prominent researchers in machine translation in Germany were able to attend the Dagstuhl seminar, and while there have decided to launch a new research program in translating spoken language in an educational context, with a particular focus on translation to German (a morphologically rich language), by submitting a Paketantrag to the DFG. The work will be carried out with a view toward creating a DFG Schwerpunktprogramm focusing on Natural Language Processing for German after the successful completion of the work in the Paketantrag. The researchers are Fraser, van Genabith, Ney, Riezler, Uszkoreit, and they are joined by Alex Waibel (who was invited to the seminar but unable to attend due to scheduling conflicts). This new funding effort would not have been possible without the possibility to meet at Dagstuhl several times to find common ground and determine an overall strategy.

In short, we were very happy with the discussions, work and impact of the Dagstuhl seminar on translation to morphologically rich languages. We plan to apply to hold a second meeting at Dagstuhl in the summer of 2016 on the same topic.

Finally, we would like to once again thank the staff of Dagstuhl for facilitating these unique scientific discussions which we are confident will have a strong impact on future research on the important problem of statistical techniques for translation to morphologically rich languages.

## 2 Table of Contents

### Executive Summary

*Alexander Fraser, Kevin Knight, Philipp Koehn, Helmut Schmid, and Hans Uszkoreit* 2

### Keynote Talks

Welcome Note, Challenges, Organizational Issues <i>Alexander Fraser</i> . . . . .	8
Morphology and Machine Translation <i>Philipp Koehn</i> . . . . .	8
Morphological Knowledge in Machine Translation <i>Kristina Toutanova</i> . . . . .	9
Explaining Data with Morphology <i>Kevin Knight</i> . . . . .	9

### Topic-Focused Talks

Modeling morphemes vs. modeling words and smoothing with morphemes <i>Chris Dyer</i> . . . . .	9
Modeling Morphology in SMT: Arabic as Example <i>Nizar Habash</i> . . . . .	10
Finding the Best Spot for Morphological Explosion <i>Jan Hajič and Ondrej Bojar</i> . . . . .	10
Language Modeling for Morphologically Rich Languages <i>Katrin Kirchoff</i> . . . . .	10
Towards a Universal Grammar for NLP? <i>Joakim Nivre and Slav Petrov</i> . . . . .	10
Synthetic Training Data <i>Joerg Tiedemann</i> . . . . .	11
(Mostly) Unsupervised Induction of Morphology for SMT <i>Sami Virpioja</i> . . . . .	11
Discourse and semantics in SMT, with attention to MRLs <i>Bonnie Webber</i> . . . . .	11
Syntactic SMT for Morphologically Rich Languages <i>Phil Williams and Philipp Koehn</i> . . . . .	12

### Working Groups

Modeling Inflectional Morphology in Statistical MT Targeting Morphologically Rich Languages <i>Nizar Habash</i> . . . . .	12
Language Modeling <i>Katrin Kirchoff</i> . . . . .	12
Semantics and SMT <i>Kevin Knight</i> . . . . .	13


Differences Between Dependency Parses and AMR	
<i>Kevin Knight</i> . . . . .	13
Universal Annotation	
<i>Joakim Nivre</i> . . . . .	13
Unsupervised Morphology	
<i>Sami Virpioja</i> . . . . .	14
Discourse	
<i>Bonnie Webber</i> . . . . .	14
Syntactic SMT	
<i>Phil Williams</i> . . . . .	14
MT Evaluation and Morphologically Rich Languages	
<i>Dekai Wu and Alon Lavie</i> . . . . .	15
<b>Participants</b> . . . . .	16



## 3 Keynote Talks

### 3.1 Welcome Note, Challenges, Organizational Issues

*Alexander Fraser (LMU München, DE)*

License  Creative Commons BY 3.0 Unported license  
© Alexander Fraser

While statistical techniques for machine translation have made significant progress in the last 20 years, results for translating to morphologically rich languages are still mixed versus previous generation rule-based systems. In particular the community working on this problem has not yet achieved coherence and as a result resources and tools can be difficult to obtain and results are sometimes not replicable. We briefly discuss these challenges to the community and present the overall organization of the seminar.

### 3.2 Morphology and Machine Translation

*Philipp Koehn (University of Edinburgh, GB)*

License  Creative Commons BY 3.0 Unported license  
© Philipp Koehn

Many aspects of translation can be best explained on a morphological, syntactic, or semantic level. Having such information available to the translation model allows the direct modeling of these aspects. For instance: reordering at the sentence level is mostly driven by general syntactic principles, local agreement constraints show up in morphology, etc.

Numerous attempts have been made to add richer information to statistical machine translation models. Most of these focus on the pre-processing of the input to the statistical system, or the post-processing of its output.


Rich morphology often poses a challenge to statistical machine translation, since a multitude of word forms derived from the same lemma fragment the data and lead to sparse data problems. If the input language is morphologically richer than the output language, it helps to stem or segment the input in a pre-processing step, before passing it on to the translation system.

One example to illustrate the short-comings of the traditional surface word approach in statistical machine translation is the poor handling of morphology. Each word form is treated as a token in itself. This means that the translation model treats, say, the word *house* completely independent of the word *houses*. Any instance of *house* in the training data does not add any knowledge to the translation of *houses*. In the extreme case, while the translation of *house* may be known to the model, the word *houses* may be unknown and the system will not be able to translate it. While this problem does not show up as strongly in English — due to the very limited morphological inflection in English — it does constitute a significant problem for morphologically rich languages such as Arabic, German, Czech, etc. Thus, it may be preferable to model translation between morphologically rich languages on the level of lemmas, and thus pooling the evidence for different word forms that derive from a common lemma.

The talk will discuss these issues in the handling of morphology, syntax and discourse. The discussion will have a particular focus on morphologically rich languages.

### 3.3 Morphological Knowledge in Machine Translation


*Kristina Toutanova (Microsoft Research – Redmond, US)*

License  Creative Commons BY 3.0 Unported license  
© Kristina Toutanova

Integrating morphological knowledge into statistical machine translation is an important challenge. This talk surveys the state-of-the-art and highlights several important findings. Unsupervised morphology is useful in MT. Pre-processing and redefining the basic units used to translate can be very effective. Factored Models generalize translation rules and incorporate more information locally. Feature-rich models for generation into morphologically rich languages improve quality. New features in standard decoders targeted at agreement and sparsity reduction increase translation quality.

### 3.4 Explaining Data with Morphology

*Kevin Knight (University of Southern California – Marina del Rey, US)*

License  Creative Commons BY 3.0 Unported license  
© Kevin Knight

We touch on morphological problems in string-based MT, syntax-based MT, and semantics-based MT. For string-based MT, we review the analysis-transfer-synthesis approach developed at IBM in 1992 by Brown and colleagues. For syntax-based MT, we suggest directions in morpho-syntax models that predict character sequences for morphologically-rich languages. Finally, we introduce an Abstract Meaning Representation for meaning-based translations.

## 4 Topic-Focused Talks

### 4.1 Modeling morphemes vs. modeling words and smoothing with morphemes


*Chris Dyer (Carnegie Mellon University, US)*

License  Creative Commons BY 3.0 Unported license  
© Chris Dyer

The talk argues for using a dual mechanism for modeling morphology. Arguments are drawn from both the psycholinguistic literature and from the state of the art natural language models for language modeling, word alignment and translation. The main argument presented in the talk is that frequent phenomena should be memorized (without generalization, which is computationally expensive), while less frequent phenomena (e.g., the plural of an infrequent noun), should be modeled using rules, resulting in high coverage. We propose the usage of hierarchical models to achieve this, and discuss possibilities for the lexicon representation.

## 4.2 Modeling Morphology in SMT: Arabic as Example


*Nizar Habash (Columbia University, US)*

License  Creative Commons BY 3.0 Unported license  
© Nizar Habash

We present some of the challenges in modeling morphology in statistical machine translation (SMT) using Arabic, a morphologically rich language. We discuss features of Arabic orthography, morphology, and morphosyntactic agreement to highlight the need for deep morphological representations in SMT.

## 4.3 Finding the Best Spot for Morphological Explosion

*Jan Hajič (Charles University – Prague, CZ) and Ondrej Bojar (Charles University – Prague, CZ)*

License  Creative Commons BY 3.0 Unported license  
© Jan Hajič and Ondrej Bojar

How should inflection (e.g., case) prediction be integrated into the decoder? We propose a discussion focusing on different ways to model inflection in the translation model and integrate this model in decoding, rather than using pre- and post-processing techniques. The slides focus on successes and failures of English to Czech MT.

## 4.4 Language Modeling for Morphologically Rich Languages


*Katrin Kirchhoff (University of Washington – Seattle, US)*

License  Creative Commons BY 3.0 Unported license  
© Katrin Kirchhoff

This talk will survey the state-of-the-art in language modeling for morphologically rich languages, particularly as applied to statistical machine translation. There is little previous work on language modeling for morphologically rich languages in statistical machine translation. There has been no systematic comparison of models. Many models simply haven't been tried yet. In work that has been tried, a better evaluation environment is needed, and in particular evaluation should just focus on the language models.

## 4.5 Towards a Universal Grammar for NLP?

*Joakim Nivre (Uppsala University, SE) and Slav Petrov (Google – New York, US)*

License  Creative Commons BY 3.0 Unported license  
© Joakim Nivre and Slav Petrov

There have been several recent initiatives in the parsing community to build treebanks with annotation that is consistent across typologically different languages. Are these resources relevant for machine translation? What needs to be added to make them (more) useful? We propose to come up with a proposal useful for a variety of purposes, including: studying

the way languages encode information, developing better models for translation, generation, parsing, making integration with other analysis and into end-applications easier, supporting cross-linguistic comparison and evaluation, and facilitating annotation of new languages.

## 4.6 Synthetic Training Data

*Joerg Tiedemann (Uppsala University, SE)*

**License** © Creative Commons BY 3.0 Unported license  
© Joerg Tiedemann

Synthetic training data and pivot languages can be used to overcome data sparseness when translating from and to morphologically rich languages. This talk will outline already studied approaches and propose new lines of work.

## 4.7 (Mostly) Unsupervised Induction of Morphology for SMT

*Sami Virpioja (Aalto University, FI)*

**License** © Creative Commons BY 3.0 Unported license  
© Sami Virpioja

We present a survey of unsupervised induction of morphology for SMT. We then describe AllomorfeSSor, which extends the unsupervised morpheme segmentation method MorfeSSor to account for the linguistic phenomenon of allomorphy, where one morpheme has several different surface forms. The method discovers common base forms for allomorphs from an unannotated corpus by finding small modifications, called mutations, for them.

## 4.8 Discourse and semantics in SMT, with attention to MRLs

*Bonnie Webber (University of Edinburgh, GB)*

**License** © Creative Commons BY 3.0 Unported license  
© Bonnie Webber


Aspects of semantic meaning and their discourse-licensed encoding in sentences can make a difference to accurate, fluent translation. Handling these aspects poses challenges to MT, especially when translating into MRLs.

Negation is one such aspect: It can be realized as a separate token, or as a morpheme attached to some root, or as an element that is itself inflected with additional information. Key to its meaning is its scope (the part of a sentence whose meaning is negated). Negation related errors in MT include: incorrectly dropping negation, incorrectly duplicating negation, and inserting negation where it will have the wrong scope.

Aspects of meaning associated with discourse itself are referring forms, semantic and pragmatic relations between sentences (and/or clauses) and information structure. Problematic for MT is the fact that these can appear in highly reduced forms (even zero) because they are obvious from the discourse context. Yet languages differ in what has to be realized explicitly, and human translators may differ in what they choose to make explicit as either lexical items, morphology or both. Dealing with information that is explicit in one language (or one half of a training pair), while implicit in the other is a particular challenge for MT.

## 4.9 Syntactic SMT for Morphologically Rich Languages

*Phil Williams (University of Edinburgh, GB) and Philipp Koehn (University of Edinburgh, GB)*

License  Creative Commons BY 3.0 Unported license  
© Phil Williams and Philipp Koehn

Languages with rich inflectional morphology pose a difficult challenge for statistical machine translation. To address the problem of morphologically inconsistent output, we add unification-based constraints to the target-side of a string-to-tree model. By integrating constraint evaluation into the decoding process, implausible hypotheses can be penalised or filtered out during search. We use a simple heuristic process to extract agreement constraints for German and test our approach on an English-German system trained on WMT data, achieving a small improvement in translation accuracy.

## 5 Working Groups

### 5.1 Modeling Inflectional Morphology in Statistical MT Targeting Morphologically Rich Languages

*Nizar Habash (Columbia University, US)*

License  Creative Commons BY 3.0 Unported license  
© Nizar Habash

The sparsity induced by target-language (TL) inflectional morphology is a fundamental challenge when translating to morphologically rich languages (MRLs). This presentation consists of three parts. First, we present a high level analysis of the various sources of TL inflectional morphology when translating from a variety of poor to MRLs. The presented analysis is supported with examples from a variety of language pairs. Second, we present a unifying description of some of the most commonly used techniques in the field for modeling morphology in statistical machine translation. We observe that one of the most elegant techniques for modeling translation of morphology has a problem with its very large search space. Much of the research on this topic is about strategies for pruning the size of the search space. Finally, we present some general insights and specific suggestions for further research directions. In particular, we think the direction of conditioning inflectional modeling using source and target language features during decoding is likely to address some of the limitations of the current state of the art.

### 5.2 Language Modeling

*Katrin Kirchhoff (University of Washington – Seattle, US)*

License  Creative Commons BY 3.0 Unported license  
© Katrin Kirchhoff

Our goal is to determine the state of the art in language modeling for SMT or MRLs. Which approaches have been tried? For which languages? Which ones work best? What are open issues/problems to be solved, in terms of modeling approaches, evaluation, resources (training

data)? We have compiled a bibliography of relevant papers on language modeling for MRLs. This will be made public on the Wiki. We discussed in detail those approaches actually used in SMT. We identified gaps/interesting problems. Finally, as a group we are writing a position paper which discusses previous and current research and proposes new directions which should be addressed.

### 5.3 Semantics and SMT

*Kevin Knight (University of Southern California – Marina del Rey, US)*

License  Creative Commons BY 3.0 Unported license  
© Kevin Knight

The Semantics and morphology group created a large number of potential questions to pursue, then discussed three of these questions in depth. The questions were: (1) How can we align strings and Abstract Meaning Representations at the token level, (2) How can semantic role labeling and other semantic features improve statistical machine translation, and (3) How can we use powerful syntax translation models to align bilingual text?

### 5.4 Differences Between Dependency Parses and AMR

*Kevin Knight (University of Southern California – Marina del Rey, US)*

License  Creative Commons BY 3.0 Unported license  
© Kevin Knight

We discuss some differences between dependencies and Abstract Meaning Representation. The talk will focus on examples motivating the need for a representation beyond dependencies, and introduce the aspects of Abstract Meaning Representation which dependencies cannot model.

### 5.5 Universal Annotation

*Joakim Nivre (Uppsala University, SE)*

License  Creative Commons BY 3.0 Unported license  
© Joakim Nivre

We propose a scheme for universal annotation useful for morphologically rich languages, based on dependencies with functional leaves. It involves two stages of tokenization/segmentation and a dual annotation of these levels. It handles rich morphology: tags, features and lemmas. By working with two levels of annotation we are able to obtain the power of preprocessing approaches without losing information.

## 5.6 Unsupervised Morphology

*Sami Virpioja (Aalto University, FI)*

License  Creative Commons BY 3.0 Unported license  
© Sami Virpioja

The challenge we have addressed in this group was to go beyond current approaches which view unsupervised morphology as inducing only a segmentation. In particular, we discussed tailoring unsupervised morphological analysis for MT and alignment. We also discussed new approaches to morphologically aware evaluation and proposed a new model incorporating both segmentation and morphology.

## 5.7 Discourse

*Bonnie Webber (University of Edinburgh, GB)*

License  Creative Commons BY 3.0 Unported license  
© Bonnie Webber

We discussed important issues of discourse with respect to translation. Our discussion was driven by this set of questions:

1. For ensuring register-level consistency in target texts, what is known about register and morphology, register and the lexicon, register and syntax?
2. What aspects of sentences need to persist throughout a discourse (to permit translation)?
3. What aspects of discourse are encoded in morphology?
4. How is scope encoded in non-configurational languages?
5. What are good test sets to use for evaluating aspects of semantics and discourse in translation?
6. What sort of “morphological divergences” occur across languages that express the same features overtly?
7. What aspects of semantics are required to be overt in some languages but not in others? (eg., evidentiality?)
8. What aspects of multiclausal (discourse) relations have been captured with syntactic transformations?
9. How to choose articles (eg., def vs indef vs generic) when you have to generate them?
10. How to tackle decoding problem when you have discourse-wide features?
11. If you have consistency, what is the consistency over?

We created a record of our discussion in the Wiki together with a list of papers and data resources for further discussion.

## 5.8 Syntactic SMT

*Phil Williams (University of Edinburgh, GB)*

License  Creative Commons BY 3.0 Unported license  
© Phil Williams

Our group attacked the following goals. We documented the easy win scenarios given the ability to enforce agreement with a focus on linguistic phenomena in languages other than

German. We focused on inflectional languages (e.g. Czech, Russian, Scandinavian and Romance languages) in this discussion. To begin our work, we documented the absolute essentials of morphosyntax in our Wiki page. After carrying out the main discussion, we discussed a number of more difficult problems we expect to encounter with highly-morphological target languages.

## 5.9 MT Evaluation and Morphologically Rich Languages

*Dekai Wu (HKUST – Hong Kong, HK) and Alon Lavie (Carnegie Mellon University, US)*

License © Creative Commons BY 3.0 Unported license  
© Dekai Wu and Alon Lavie

The most commonly used MT evaluation metrics to date, both human and automatic, have done little to address the issues in morphologically rich languages. Both inflectional morphology (as found in languages such as German, Arabic, Finnish, or Czech) and derivational morphology / compounding (as found in all languages but even more acutely problematic in languages such as Chinese, Finnish, or Turkish) cause simple n-gram oriented metrics to significantly underestimate translation accuracy. These issues impact SMT training and tuning when automatic metrics are used as the objective functions. Emerging work on MT evaluation metrics, incorporating explicit morphological components as in METEOR, and/or explicit semantic parsing components as in MEANT, represent strategies to abstract away from surface form n-grams so as to better handle morphological variation. We discuss and analyze key open questions, leading to a roadmap for research to address the deficiencies of MT evaluation metrics for morphologically rich languages.



## Participants

- Arianna Bisazza  
University of Amsterdam, NL
- Fabienne Braune  
Universität München, DE
- Fabienne Cap  
Universität München, DE
- Marine Carpuat  
NRC – Ottawa, CA
- David Chiang  
University of Southern California  
– Marina del Rey, US
- Ann Clifton  
Simon Fraser University –  
Burnaby, CA
- Hal Daumé III  
University of Maryland, US
- Gideon Maillette de Buy  
Wenniger  
University of Amsterdam, NL
- Chris Dyer  
Carnegie Mellon University, US
- Andreas Eisele  
European Commission  
Luxembourg, LU
- Richard Farkas  
University of Szeged, HU
- Marcello Federico  
Bruno Kessler Foundation –  
Trento, IT
- Mark Fishel  
Universität Zürich, CH
- Anette Frank  
Universität Heidelberg, DE
- Alexander M. Fraser  
LMU München, DE
- Spence Green  
Stanford University, US
- Nizar Habash  
Columbia University, US
- Jan Hajič  
Charles University – Prague, CZ
- Katrin Kirchhoff  
University of Washington –  
Seattle, US
- Kevin Knight  
Univ. of Southern California –  
Marina del Rey, US
- Philipp Koehn  
University of Edinburgh, GB
- Jonas Kuhn  
Universität Stuttgart, DE
- Alon Lavie  
Carnegie Mellon University, US
- Krister Linden  
University of Helsinki, FI
- Andreas Maletti  
Universität Stuttgart, DE
- Maria Nadejde  
University of Edinburgh, GB
- Preslav Nakov  
QCRI – Doha (Qatar), QA
- Hermann Ney  
RWTH Aachen, DE
- Joakim Nivre  
Uppsala University, SE
- Slav Petrov  
Google – New York, US
- Maja Popovic  
DFKI – Berlin, DE
- Anita Ramm  
Universität München, DE
- Stefan Riezler  
Universität Heidelberg, DE
- Hassan Sajjad  
QCRI – Doha (Qatar), QA
- Helmut Schmid  
LMU München, DE
- Hinrich Schütze  
LMU München, DE
- Khalil Sima'an  
University of Amsterdam, NL
- Sara Stymne  
Uppsala University, SE
- Jörg Tiedemann  
Uppsala University, SE
- Kristina Toutanova  
Microsoft Res. – Redmond, US
- Hans Uszkoreit  
Universität des Saarlandes, DE
- Josef van Genabith  
Dublin City University, IE
- Sami Virpioja  
Aalto University, FI
- Stephan Vogel  
QCRI – Doha (Qatar), QA
- Martin Volk  
Universität Zürich, CH
- Bonnie Webber  
University of Edinburgh, GB
- Marion Weller  
Universität München, DE
- Phil Williams  
University of Edinburgh, GB
- Shuly Wintner  
Haifa University, IL
- Dekai Wu  
HKUST – Hong Kong, HK
- François Yvon  
University Paris Sud, FR

