Linköping Studies in Science and Technology.

Dissertations. No. 1451

Text Harmonization Strategies for Phrase-Based Statistical Machine Translation

Sara Stymne

Department of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden

Linköping 2012

ISBN: 978-91-7519-887-3

ISSN: 0345–7524

URL: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-76766

Cover, Photography: Jalal Maleki; Layout: Martin Pettersson; Architects and builders: Amy Rankin, Christian Smith, Jody Foo, Johan Blomkvist, Jonas Rybing, Lisa Malmberg, Robin Keskisärkkä, and Sara Stymne

Typeset using LaTeX

Printed by LiU-Tryck, Linköping 2012

Abstract

In this thesis I aim to improve phrase-based statistical machine translation (PBSMT) in a number of ways by the use of text harmonization strategies. PBSMT systems are built by training statistical models on large corpora of human translations. This architecture generally performs well for languages with similar structure. If the languages are different for example with respect to word order or morphological complexity, however, the standard methods do not tend to work well. I address this problem through text harmonization, by making texts more similar before training and applying a PBSMT system.

I investigate how text harmonization can be used to improve PBSMT with a focus on four areas: compounding, definiteness, word order, and unknown words. For the first three areas, the focus is on linguistic differences between languages, which I address by applying transformation rules, using either rule-based or machine learning-based techniques, to the source or target data. For the last area, unknown words, I harmonize the translation input to the training data by replacing unknown words with known alternatives.

I show that translation into languages with closed compounds can be improved by splitting and merging compounds. I develop new merging algorithms that outperform previously suggested algorithms and show how part-of-speech tags can be used to improve the order of compound parts. Scandinavian definite noun phrases are identified as a problem for PBSMT in translation into Scandinavian languages and I propose a preprocessing approach that addresses this problem and gives large improvements over a baseline. Several previous proposals for how to handle differences in reordering exist; I propose two types of extensions, iterating reordering and word alignment and using automatically induced word classes, which allow these methods to be used for less-resourced languages. Finally I identify several ways of replacing unknown words in the translation input, most notably a spell checking-inspired algorithm, which can be trained using character-based PBSMT techniques.

Overall I present several approaches for extending PBSMT by the use of pre- and postprocessing techniques for text harmonization, and show experimentally that these methods work. Text harmonization methods are an efficient way to improve statistical machine translation within the phrase-based approach, without resorting to more complex models.

Populärvetenskaplig sammanfattning

Textharmonisering, en metod för att förbättra maskinöversättning

Maskinöversättning, det vill säga automatisk översättning mellan naturliga språk som engelska och svenska, spelar en allt större roll när världen blir alltmer internationell. Stora organisationer som EU och FN och multinationella företag har ett enormt översättningsbehov. Här kan maskinöversättning vara till hjälp, framförallt som stöd för översättare. Men även privatpersoner kan ha nytta av maskinöversättning. Till exempel kan vi nu tillgodogöra oss stora delar av innehållet på en webbsida även om vi inte förstår språket den ursprungligen är skriven på.

I den här avhandlingen presenterar jag mitt arbete kring hur statistisk maskinöversättning kan förbättras genom att använda metoder för textharmonisering. Statistisk maskinöversättning bygger på att man skapar statistiska modeller för hur översättningar ser ut genom att utgå från stora samlingar av mänskliga översättningar, så kallade parallella korpusar. De metoder som används vid statistisk översättning fungerar bäst om språken har liknande struktur, till exempel inte har stora skillnader i ordföljd. Jag har angripit detta problem genom textharmonisering, det vill säga att förändra texterna på det ena språket på något sätt så att de får en struktur liknande det andra språket, och därigenom förbättra översättningsresultaten.

Jag har visat att översättningar från engelska till svenska och tyska kan förbättras genom att man delar upp sammansättningar på svenska och tyska, så att deras struktur mer liknar engelskans, där sammansättningar i regel skrivs som flera enskilda ord. Framförallt har jag utvecklat nya metoder för hur sammansättningsdelarna ska hamna i rätt följd och för hur de ska sättas ihop igen efter översättningsprocessen. För översättning till skandinaviska språk har jag designat en algoritm som gör engelska fraser som *the dog* mer lika skandinaviska, *hunden*, genom att till exempel plocka bort onödiga bestämda artiklar, vilket leder till förbättrade översättningar. För språk med olika ordföljd har jag vidareutvecklat befintliga algoritmer, framförallt genom att använda ordklasser som tas fram automatiskt från text, vilket gör att man kan använda dessa algoritmer för språk med få språkresurser, som till exempel haitisk kreol. Slutligen har jag visat hur man kan ersätta ord som är okända för maskinöversättningssystemet med ord som systemet känner till, bland annat baserat på metoder för stavningskontroll.

I avhandlingen beskrivs dessa och flera andra metoder för textharmonisering, och jag visar hur de kan användas för att på ett effektivt sätt förbättra statistisk maskinöversättning.

Acknowledgements

I could not have written this thesis without the support of others, whom I cannot thank enough.

First of all I would like to thank Lars Ahrenberg, my primary supervisor. You have been an enormous help in letting me evolve as a researcher, letting me explore new avenues and change my focus when I wanted to, but also helping me to stay on track and to remember the big picture when I got too buried in details. Another thank you to Joakim Nivre, my secondary supervisor, for many valuable discussions about algorithms, strategies, and many other things, and for all your feedback on different texts.

Thank you also to all the members of CILTLab, formerly known as NLPLab, for creating a nice working atmosphere with so many opportunities for both formal and informal discussions about research and many other things. Thank you to Jody Foo, Arne Jönsson, Jalal Maleki, Magnus Merkel, and Christian Smith for good collaborations on research and teaching, and for fun conference travel. Thank you to Nils Dahlbäck and Annika Silvervarg for welcoming me into the SweCog community. A special thank you to Maria Holmqvist; these years would not have been the same without you. Thank you for helping me understand how to interpret things when I started out, for introducing me to SMT, which completely changed my research focus, for always having a moment when a new idea popped up in my head or when I couldn't fix a bug, and for all the fun times we've had.

Thanks also to all my fellow PhD students, past and present, and others who joined in on coffee breaks, lunches, pub nights, board game nights, and the like. Maria, Jody, Christian, Mattias, Camilla, Fabian, Johan, Amy, Lisa, Magnus, Jonas, Robin, Anna, Susanna, Jiri, Ola, Anders, and Björn, you really helped me to enjoy life at IDA. An extra thank you to those of you who helped me build Babel's tower. And a very big thank you to Jalal for photographing it so well.

A big thank you to Santa Anna, especially to Sture Hägglund, and to GSLT for financing my PhD. I would also like to thank all GSLT members for all the great times we had taking courses, discussing research, and sharing all those nerdy jokes. I'm privileged to have been part of such a great research community.

I also would like to thank everyone in the CLT group, or whatever your name is now, at XRCE in Grenoble, where I spent a total of nine months, doing parts of the research in this thesis. Thank you all for making me feel so welcome, and for really letting me be part of the group. Thank you to Nicola Cancedda for making it possible for me to come there twice and for great collaborations.

Another big thank you to all the technical and administrative staff at IDA. I would especially like to thank Björn and Nisse for your very quick response last time my computer broke down, Arne for giving me some extra quota at the end of writing, Rikard for helping me run experiments on the cluster, Peter J. for all the times you quickly solved my problems, Rolf for getting me a new monitor and for always being so cheerful, Anne for all your help these last few months, guiding me in the thesis jungle, and Lise for all your help throughout the years.

Thank you to Jan Niehues for sharing his reordering code, and to everyone who contributed to open source tools, like Moses, which were essential for my research. Thank you to Alexander Fraser for good discussions at my licentiate seminar. I would also like to thank all anonymous reviewers for their comments both on the papers in this thesis, and on other papers. Your feedback has been really valuable. I also appreciate everyone who discussed my work during conferences and workshops.

I would like to thank my family, Claes, Anita, Hanna, Petter, Uwe, Linda, and everyone else for always being supportive and LUKAS, ELIN, LEVI, and THEA for brightening up my life. I would also like to thank all my friends in Linköping, Grenoble, and elsewhere for being there, and for taking my mind off things when needed. Finally I would like to thank Johan for being you, and for being there for me.

Contents

1 Introduction 1

1.1 Text harmonization for SMT . . . 3
1.2 Contributions . . . 4
1.3 Included papers . . . 5
1.4 Outline . . . 7

2 Statistical machine translation 9

2.1 Approaches to statistical machine translation . . . 9
2.1.1 Word-based SMT . . . 9
2.1.2 Phrase-based SMT . . . 12
2.1.3 Factored SMT . . . 17
2.1.4 Tree-based SMT . . . 19
2.2 Pre- and postprocessing . . . 20
2.2.1 Preprocessing . . . 20
2.2.2 Postprocessing . . . 21
2.3 Machine translation evaluation . . . 23
2.3.1 Human evaluation . . . 23
2.3.2 Automatic evaluation . . . 24

3 Text harmonization: Focus areas and strategies 29

3.1 Compounding . . . 29
3.1.1 Compounds in English . . . 31
3.1.2 Compounds in German, Swedish, and Danish . . . 32
3.1.3 Compound processing . . . 34
3.2 Definiteness . . . 38
3.2.1 Definiteness in English and Italian . . . 39
3.2.2 Definiteness in the Scandinavian languages . . . 40
3.2.3 Definiteness processing . . . 42
3.3 Word order . . . 43
3.3.1 Word order in English and German . . . 44
3.3.2 Word order in Haitian Creole and English . . . 46
3.3.3 Reordering . . . 47
3.4 Unknown words . . . 49
3.4.1 Non-standard orthography . . . 50
3.4.2 Online OOV-handling . . . 52

4 **Experiments and results** **55**
4.1 Research approach 55
4.2 Tools and corpora 57
4.3 Paper summaries 59
4.3.1 Paper 1 61
4.3.2 Paper 2 62
4.3.3 Paper 3 63
4.3.4 Paper 4 64
4.3.5 Paper 5 64
4.3.6 Paper 6 65
4.3.7 Paper 7 66

5 **Conclusions** **69**

Bibliography **73**

Paper 1: Generation of compound words in statistical machine translation into compounding languages **97**

Paper 2: Definite noun phrases in statistical machine translation into Danish **135**

Paper 3: Definite noun phrases in statistical machine translation into Scandinavian languages **143**

Paper 4: Iterative reordering and word alignment for statistical MT **153**

Paper 5: Clustered word classes for preordering in statistical machine translation **159**

Paper 6: Vs and OOVs: Two problems for translation between German and English **169**

Paper 7: Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation **177**

1 Introduction

Translation is the task of transferring a text, written in a source language, into another language, a target language. In order to translate a sentence properly a human needs knowledge of both languages, to understand the source text, and to be able to produce a well-formed target language text. In addition, knowledge about the subject matter, the purpose of the text, and the intended readers are prerequisites for a good translation.

Compared to human translation, machine translation (MT), that is, automatic translation by computers, is even more challenging. To code all types of human knowledge into a machine would be very hard, and is not attempted by any MT systems. However, the rule-based approach to MT is to code a well-defined subset of human language. It is based on rules, often hand-written, that typically analyze the source sentence, transfer it to a target-side representation, and generate target language. Usually some type of syntactic analysis is performed, possibly with some semantics, but there are also systems that use more shallow representations. An alternative type of approach to machine translation is the empirical approach where existing human translations are the basis of the translation process. In this thesis the focus will be on statistical MT (SMT), where statistical models are trained automatically from parallel corpora of human translations. The translation strategy adopted in this thesis is phrase-based statistical machine translation (PBSMT), where the translation unit is the phrase, a sequence of words.

PBSMT is a successful approach to MT and it is currently the dominant approach in research on MT. PBSMT systems have the advantage of being easy and fast to build as long as there is a suitable parallel corpus, which, however, is not always the case. The core methods are language independent; the models are trained in the same way regardless of which language pair that is treated. This is an advantage when training a new system, but it has the disadvantage of not taking advantage of any specific knowledge about a language pair, which might otherwise improve the translation process. In the standard PBSMT approach no linguistic knowledge is used at all, which can lead to ungrammatical output.

Some problems with PBSMT are exemplified in Table 1.1.[1] In the first example, for German, the translation of the verb *welcome* is missing in the

[1] In tables and examples languages will be presented with ISO639-1 language

Table 1.1: Examples of problematic PBSMT output for different language pairs, contrasted to human reference translations. The PBSMT examples are from different baseline systems from the thesis.

En source	I too would like to welcome Mr Prodi's forceful and meaningful intervention.
De PBSMT	Ich möchte auch herrn Prodis energisch und sinnvollen Beitrag.
De reference	Ich möchte meinerseits auch den klaren und substanziellen Redebeitrag von Präsident Prodi begrüßen.
En source	So much for the scientific approach.
Sv PBSMT	Så mycket för den vetenskapliga synsätt.
Sv reference	Så mycket för den vetenskapliga infallsvinkeln.
Ht source	Yon ti jès pou mwen, yon tandrès pou mwen, yon pawòl pou mwen, yon zepòl pou mwen
En PBSMT	A little jès for me, tandrès for me. A rumor for me, zepòl for me.
En reference	A gesture, some affection for me, some words for me, a shoulder for me

PBSMT output. Missing and misplaced verbs are common error types, since the German verb should appear last in the sentence in this context, as in the reference, *begrüßen*. There is also an idiomatic compound, *Redebeitrag* (speech+contribution; intervention) in the reference, which is produced as the single word *Beitrag* in the PBSMT output. In the Swedish example, there are problems with a definite noun phrase (NP), which has the wrong gender of the definite article, *den* instead of *det*, and is missing a definite suffix on the noun *synsätt(et)* ([the] approach). In the last example from Haitian Creole to English, there are three unknown Haitian Creole words that are directly transferred to the output, *jès* (gesture), *tandrès* (affection), and *zepòl* (shoulder).

As illustrated by these examples, there are many problems to address in order to improve translation quality. In general, standard PBSMT techniques tend to work better for similar languages. Issues such as differing word order or different word segmentation tend to affect PBSMT negatively. Thus, one way to improve translations is by the use of *text harmonization*, the process of making two texts more similar, which is the focus of this thesis.

codes: Danish–da, English–en, German–de, Haitian Creole–ht, Italian–it, Norwegian Bokmål–nb, and Swedish–sv.

1.1 Text harmonization for SMT

The goal of the thesis is to improve SMT by the use of text harmonization, that is, to transform one text to become more similar to another text in some respect. The term text harmonization has not been used much in relation to SMT, even though the term harmonization has been used sporadically, mostly in relation to work that addresses word order issues. I think it is a useful umbrella term for a set of commonly used methods.

Text harmonization can be performed on different texts used for SMT. One option is to make the source text more similar to the target text, or vice versa. In that case it is possible to address linguistic differences between two languages. Another option is to make the translation input more similar to the training data, for instance by the replacement of unknown words (out-of-vocabulary words, OOVs) in the translation input. A third possibility is to make the training and/or input data more similar to standardized language, which could be useful if the data is noisy.

I have focused on how text harmonization can be used to address four areas that are problematic for current PBSMT:

- Compounding
  - Issue: Compounds are closed, that is, written like single words in some languages, for instance German and Swedish, and open, written as multiple words, in other languages such as English.
  - Harmonization with regard to compounds: Splitting closed compounds, thus making the compound structure more similar to languages with open compounds.
- Definiteness
  - Issue: Definiteness in Scandinavian languages can be expressed by a definite article and/or a definite noun suffix. In other languages, such as English, only the definite article is used.
  - Harmonization with regard to definiteness: Changing the structure of noun phrases in non-Scandinavian languages, to look like Scandinavian noun phrases, by removing definite articles, and adding noun suffixes.
- Word order
  - Issue: Word order is different in natural languages. I have mainly focused on English and German between which there are several word order differences for instance regarding verb placement.

  - Harmonization with regard to word order: Creating rules that can change the word order of one language to become more similar to the other language.

- Unknown words
  - Issue: Words in the translation input that have not been seen in the training data cannot be translated by a PBSMT system.
  - Harmonization with regard to unknown words: Replacing unknown words in the translation input by known words identified by the use of techniques such as spell-checking and morphological processing.

1.2 Contributions

I have investigated how text harmonization strategies can be used to tackle different areas that are problematic to SMT. The main contributions of the thesis are the following:

- I have thoroughly investigated several areas relevant to the generation of novel compounds for translation into compounding languages.
  - I have shown how customized part-of-speech (POS) tagsets can be used in sequence models and as count features in order to improve compound formation.
  - I have developed both heuristic and machine learning algorithms for compound merging that outperform previously proposed algorithms.
- I have identified Scandinavian definite noun phrases as a problem for SMT in translation into Scandinavian languages, and suggested a preprocessing approach to overcome this problem.
- I have investigated how existing approaches to reordering can be augmented in different ways.
  - I have presented an approach where word alignment and reordering are iterated, and shown that different types of linguistically motivated rules can be learnt in different iterations.
  - I have shown that clustered word tags can successfully be used instead of standard POS-tags in a preordering approach.
- I have identified several ways to replace OOVs in the translation input, most notably a spell checking inspired algorithm, for which weights could be trained automatically by using character-based PBSMT techniques.

1.3 Included papers

This thesis contains seven articles that focus on text harmonization for PB-SMT.

Paper 1 Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2012a. Generation of compound words for statistical machine translation into compounding languages. Submitted manuscript.
In this paper we discuss compound processing for translation into German, Swedish, and Danish. The focus is on effects of compound splitting and merging strategies and on how to improve the coalescence of compound parts by the use of sequence models on customized POS-tagsets.
Most of the ideas were originally mine; the idea about merging as sequence labeling was originally Nicola Cancedda's. All ideas were developed in discussion between the authors. I was responsible for the implementation and evaluation. I wrote most of the paper, aided by the two co-authors.

Paper 2 Sara Stymne. 2009c. Definite noun phrases in statistical machine translation into Danish. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 4–9. Odense, Denmark.
In this paper I present a preprocessing strategy for improving the treatment of definite noun phrases for translation into Danish.

Paper 3 Sara Stymne. 2011a. Definite noun phrases in statistical machine translation into Scandinavian languages. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 289–296. Leuven, Belgium.
This paper is an extension of Paper 2, where the preprocessing strategies for definite NPs are adjusted for and applied to other language pairs than English–Danish.

Paper 4 Sara Stymne. 2011b. Iterative reordering and word alignment for statistical MT. In *Proceedings of the 18th Nordic Conference on Computational Linguistics (NODALIDA'11)*, pages 315–318. Riga, Latvia.
In this paper I investigate how preprocessing strategies for learning reordering rules and word alignment can be iterated, in order to find new types of rules and improve alignments.

Paper 5 Sara Stymne. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 28–34. Avignon, France.
In this paper I show that clustered word classes can be used in place of standard POS-tags for learning reordering rules.

Table 1.2: Overview of the work on the four focus areas, the relation to the papers, and the languages used.

Focus area	Paper	Languages
Compounding	1 (6)	en⇒{de,sv,da}
Definiteness	2–3	en⇒{sv,da,nb}, it⇒da
Word order	4–5 (6)	en⇒de, ht⇒en
Unknown words	6–7	en⇔de, ht⇒en

Paper 6 Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and OOVs: Two problems for translation between German and English. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 183–188. Uppsala, Sweden.
In this paper we explore how knowledge-light preprocessing strategies can be used to replace OOVs in the translation input, by the use of techniques like stemming. The paper also contains experiments on reordering for alignment and compound processing.
The work on OOVs and compounding were based on my ideas and implemented by me. Maria Holmqvist did most of the work on reordering for alignment. The paper was jointly written by all three authors.

Paper 7 Sara Stymne. 2011c. Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 470–477. Edinburgh, Scotland.
In this paper I explore the use of spell-checking techniques for the replacement of OOVs in the translation input. I also investigate how a small amount of data cleaned by humans can be used to train a model for cleaning a noisy corpus.

These papers will be referred to as Papers 1–7 throughout this thesis. Table 1.2 gives an overview of how these papers relate to the focus areas of this thesis and also shows which language pairs that were used in the experiments.

This PhD thesis is an extension of my licentiate thesis *Compound processing for phrase-based statistical machine translation* (Stymne, 2009b), which discussed compound processing. The PhD thesis contains new work on compound processing, and also work related to definiteness, reordering and OOV processing. The introductory chapters of the PhD thesis partially overlap with the licentiate thesis.

1.4 Outline

Chapter 1 is a brief introduction to the subject matter, and also contains the contributions of the thesis. Chapters 2 and 3 are background chapters. Chapter 2 gives an overview of SMT and chapter 3 gives on overview of the focus areas of the thesis, and of previous work on text harmonization. Chapter 4 contains an overview of the experiments in the thesis, and summaries of the papers. Chapter 5 contains the conclusions. Finally there are the seven included papers.

2 Statistical machine translation

This chapter contains three parts. First there is an overview of approaches used for statistical machine translation (SMT), with the main focus on phrase-based SMT (PBSMT). Then there is a brief review of pre- and post-processing techniques for SMT, followed by a description of methods for machine translation evaluation.

2.1 Approaches to statistical machine translation

SMT is a form of empirical machine translation. It is based on statistical models that can be automatically trained based on large parallel corpora of human translations. This can be contrasted to rule-based methods, where knowledge is generally coded into resources such as grammars and transfer rules by human experts. In this section I first review word-based SMT, which is the oldest type of SMT. I then go on to introduce phrase-based SMT, which is the framework used in this thesis. I finally briefly review tree-based SMT.

2.1.1 Word-based SMT

Traditional statistical MT uses words as the translation unit and is based on the noisy channel model, shown using Bayes' rule in Equation 2.1,[1] where we want to find the probability of a target sentence, T, given a source sentence, S. To find the best translation, $\hat{T}$, Equation 2.1 can be re-written as 2.2, where the denominator, $P(S)$, is removed, since the probability of the source sentence is constant. $P(S|T)$, is given by a translation model and $P(T)$ is given by a language model. In addition, to find the best translation a decoder is needed, which given a source sentence, S, produces the most

[1] The language independent notation S for source language and T for target language is used in this thesis, not the commonly used notation of E for English and F for French or foreign.

probable target sentence T, or possibly an n-best list of the most probable translations.

$$P(T|S) = \frac{P(S|T) \cdot P(T)}{P(S)} \tag{2.1}$$

$$\hat{T} = \arg\max_{T} P(S|T) \cdot P(T) \tag{2.2}$$

Language model

The language model, $P(T)$, accounts for the fluency of the translation, it gives a probability for a sequence of words being a target sentence. It is common to use n-gram-based language models that build on the Markov assumption that the probability for each word can be based on the n previous words. The probability for a sentence is calculated as the product of the probability of each word, given a history of $n-1$ previous words. In a bigram model, where $n = 2$, this means that the probability for each word is only conditioned on the previous word, and the probability for the sentence *The old man sleeps.* would be calculated as in Equation 2.3, where BOS and EOS are beginning and end of sentence markers.

$$\begin{aligned} P(\text{The old man sleeps .}) \;=\; & P(\text{The}|\text{BOS}) \cdot P(\text{old}|\text{The}) \cdot P(\text{man}|\text{old}) \cdot \\ & P(\text{sleeps}|\text{man}) \cdot P(.|\text{sleeps}) \cdot P(\text{EOS}|.) \end{aligned} \tag{2.3}$$

These probabilities can be estimated from a mono-lingual corpus using maximum-likelihood estimation, as in Equation 2.4 for trigrams, where $C(w)$ is the count function for the word sequence w, for instance, $C(w_{n-1}, w_n)$ is the count of the bigram w_{n-1}, w_n in a corpus (Manning and Schütze, 1999). Even if an n-gram model is trained on a large amount of data, it will suffer from data sparseness, i.e., many n-grams will have been seen few or no times at all. This is addressed by the use of smoothing techniques, where some of the probability mass of seen events are given to unseen or rare events (see e.g., Manning and Schütze, 1999, for an overview).

$$P(w_n|w_{n-2}, w_{n-1}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})} \tag{2.4}$$

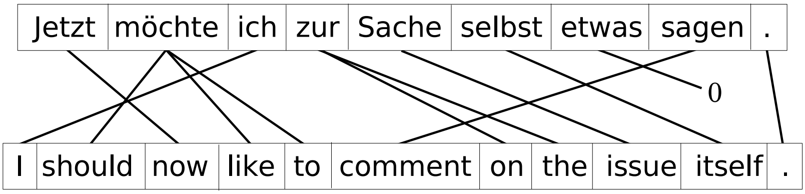

Figure 2.1: Example of a word aligned sentence

Translation model

The translation model, $P(S|T)$, accounts for the adequacy of the translation, that is, how faithful the translation is. It is normally estimated from a parallel bilingual corpus. Statistical translation models estimate the conditional probability of a target sentence given a source sentence using word alignments. In a word aligned text, words that correspond to each other are linked, as shown in Figure 2.1. Some words have no correspondences in the other language, such as *etwas* (something), which then receives a so called null link. The translation model can be calculated as the sum over all possible alignments, as in Equation 2.5, where A is an alignment.

$$P(S|T) = \sum_{A} P(S, A|T) \tag{2.5}$$

IBM researchers (Brown et al., 1993) developed a series of five increasingly more complex models that estimate translation models and word alignments from sentence-aligned text, called the IBM models. The first model only takes into account the translation of words into other words. In models 2–5 distortion is taken into account as well and in models 3–5 fertility is also added. Distortion is a measure of how target words are reordered, compared to the source. Fertility is a measure of how many source words a single target word is translated into.

The IBM models do not directly estimate the probability in Equation 2.5. A somewhat simplified equation for models 3–5, formulated by (Elming, 2008a), is shown in Equation 2.6, where i is a position of the target sentence t with length l, j is a position in the source sentence s with length m, a_j is the position of the target word that word j is aligned to, and ϕ_i is the fertility of target word i. The equation has three parts, the probability n giving the fertility for each target word, how many words each target word generates, the probability tr, that a source word form translates into a target word form, and the distortion probability d, the probability that a word form appears in a source sentence position, given the link to a target sentence position, and the length of the sentences.

$$P(S, A|T) = \prod_{i=1}^{l} n(\phi_i|t_i) \prod_{j=1}^{m} tr(s_j|t_{a_j}) \prod_{j=1}^{m} d(j|a_j, m, l) \tag{2.6}$$

To estimate these probabilities the expectation-maximization (EM) algorithm (Dempster et al., 1977) is used. The EM algorithm is an iterative method with two steps. In the expectation step expected alignment frequencies are estimated based on the current model parameters. In the maximization step the model parameters are re-estimated based on the alignment frequencies. The EM algorithm is only guaranteed to reach a local maximum, which makes it sensitive to the initial estimation of the model parameters. Therefore, the models are often run in sequence, where the result of the lower models is used to initialize the next model. IBM model 2 is often replaced by a Hidden Markov model (HMM) described by Vogel et al. (1996). All these models are asymmetric and create one-to-many alignments, i.e., one word in the target text can be aligned to many source words, but each source word can only be aligned to one target word.

2.1.2 Phrase-based SMT

In phrase-based SMT (Och et al., 1999; Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003), the unit of translation is not a single word but a phrase. A phrase in this context is a sequence of words, not necessarily a linguistically motivated phrase. Figure 2.2 shows two examples of a phrase-aligned sentence, with different granularity of the phrases. In this type of model there is an n–n alignment between phrases, that is, the number of phrases is equal on both sides, but the size of the individual phrases can vary. Phrase-based SMT was first suggested by Och et al. (1999) under the name alignment template model, which was based on word classes. A similar approach to PBSMT is n-gram-based translation (Mariño et al., 2006), based on bilingual n-grams.

At the same time as phrase-based models were suggested, there was also a move from the traditional noisy channel model to log-linear models (Och and Ney, 2002), which became the standard model type used for PBSMT. In a log-linear model, the probability $P(T|S)$ is modeled by a set of M feature functions $h_m(T, S)$, where each feature function has a weight λ_m. The best sentence, $\hat{T}$, is computed as in Equation 2.7, where Z_s is a normalization constant. The feature functions include the language model and the translation model.

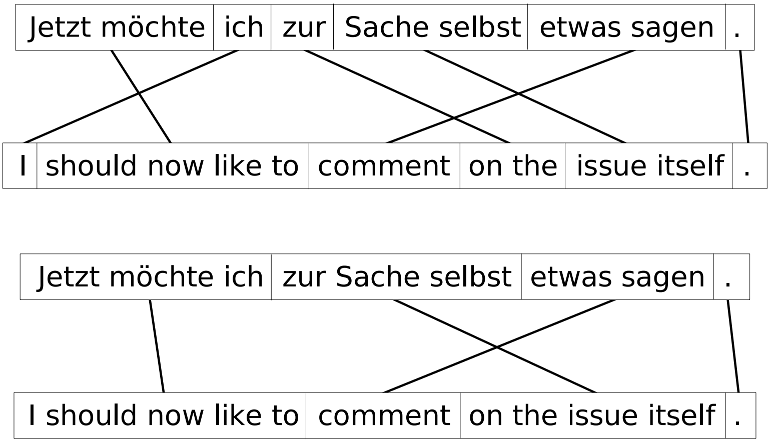

Figure 2.2: Examples of phrase-alignments with different granularity

$$\begin{aligned}\hat{T} &= \arg\max_{T} P(T|S) \\ &= \arg\max_{T} \frac{1}{Z_s} exp\left(\sum_{m=1}^{M} \lambda_m h_m(T,S)\right)\end{aligned} \tag{2.7}$$

The language model is normally the same for phrase-based as for word-based translation. The main difference from word-based models is in the translation model, which now includes probabilities for translating phrases, not only single words. An advantage of log-linear models is that it is easy to add other feature functions than just the language and translation models. It is common for instance to add more advanced distortion models, and word and phrase penalties, that can control the length of the output sentence and the tendency to choose long or short phrases.

In most phrase-based approaches the extracted phrases are contiguous, that is, non-interrupted sequences of words. Another possibility is to allow discontiguous phrase alignments, where there can be gaps in the phrases. An example of a discontiguous phrase alignment is shown in Figure 2.3, where discontiguous links are used to handle a German phrasal verb with a separated verb prefix, *fängt . . . an* (start), and an English expression, *it is best*, interrupted by the adverb *always*. Goutte et al. (2004) described an extension to standard PBSMT which allows fixed-sized gaps in the extracted phrases and Simard et al. (2005) described a decoder that used the discontiguous phrases. Galley and Manning (2010) described a decoder for

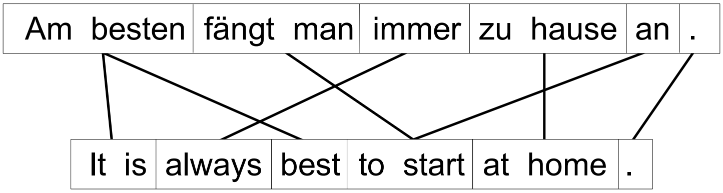

Figure 2.3: Example of phrase-alignment with discontiguous links

handling gaps of variable size, based on the alignment algorithm described in Lopez (2007).

Translation model

The translation model contains probabilities for phrase translations. A common way to construct a translation model for PBSMT, described in Koehn et al. (2003), is to start with asymmetric one-to-many word alignments in both directions, extracted based on the IBM models, which are then symmetrized into many-to-many alignments. From this alignment consistent phrases are extracted and scored. There are other possibilities, such as to estimate phrase probabilities directly from the corpus, not via word alignments (Marcu and Wong, 2002), which has, however, been shown to perform worse than word-alignment-based methods.

Symmetrization normally starts with the intersection of the two unidirectional alignments, and proceeds by adding a subset of the links from the union. Och and Ney (2000) described a refined symmetrization method, where they add alignment points from the union if they align at least one unaligned word, and are horizontal or vertical neighbors of an alignment point, or if they connect previously unaligned words. Koehn et al. (2005) described an alternative to this method, grow-diag-final-and, where diagonal neighbors are also allowed, and where unaligned points are added in a final step if they connect two previously unaligned words.

From a symmetrized alignment, Koehn et al. (2003) created a phrase alignment by collecting all phrase pairs that are consistent with the word alignment, that is, the words in a phrase pair can only be aligned with words in the same pair, not to words outside the phrase pair. The probabilities were estimated by relative frequencies, as in Equation 2.8, where (s,t) is a phrase correspondence, an alignment between two phrases.

$$\phi(s|t) = \frac{count(s,t)}{\sum_{s'} count(s',t)} \tag{2.8}$$

Koehn et al. (2003) suggested using lexical weighting as a complement to phrase probabilities. The lexical weighting is a probability that is based on the probabilities of the word alignments between individual words in a phrase pair. Both for phrase probabilities and lexical weighting, it is common to use probabilities for both translation directions, i.e., not only $P(s|t)$, but also $P(t|s)$.

Goutte et al. (2004) described a method for directly estimating discontiguous phrases from n-best lists of one-to-many alignments. A word alignment matrix was created with weights based on how many times each pair of words was aligned in the n-best lists. From this matrix a parallel partitioning of the words is created based on orthogonal non-negative matrix factorization. Each partition is a phrase pair, which can be discontiguous.

Distortion models

In PBSMT a large part of the local reordering is taken care of within phrase pairs. The phrase pairs can capture local reorderings that were seen in the training data, as in (1) where we can extract the phrase pair *Gestern erlebten wir – Yesterday, we experienced*, to account for the fact that the German subject follows the verb after an adverbial. These reorderings are, however, only local and cannot be generalized, so there is still a need to model distortion in phrase-based models.

(1) Gestern erlebten wir die Verhaftung ...
Yesterday experienced we the arrest ...

Yesterday, we experienced the arrest ...

Normally some method that restricts the possible reorderings is applied, in order to reduce the complexity of decoding from NP-hard without reordering constraints to polynomial (Zens et al., 2004). One common restriction is the IBM constraint (Berger et al., 1996), which allows unlimited reorderings, but with a limit on how far words can move, thus disallowing global reordering. A special case of the IBM constraint is when the reordering limit is set to 0, that is, no reorderings are allowed, which is called monotone decoding. Monotone decoding is efficient, but in general gives bad results. It can, however, be used together with other methods that address reordering. Another type of reordering constraint is based on inversion transduction grammars (The ITG constraint, Wu, 1997; Zens et al., 2004). It allows global reordering, but the kinds of permutations are restricted.

It is common to use a distortion penalty, a flat penalty that punishes any deviation from the source order of phrases. The distortion penalty simply adds a factor δ^n for movements over n words. The distortion penalty only takes the position of phrases into account, not the words in them. In addition it is common to impose a constraint, a distortion limit, on the maximum distance a phrase can move. This default distortion model is weak; it discourages distortion, but allows some distortion to take place if it has support from the language model.

A number of alternative distortion models, with a higher degree of discrimination of orderings have been suggested (e.g., Tillman, 2004; Koehn et al., 2005; Al-Onaizan and Papineni, 2006; Kuhn et al., 2006; Galley and Manning, 2008). Koehn et al. (2005) described a lexicalized reordering model, which for each phrase learns how likely it is to follow the previous phrase (monotone), swap places with the previous phrase (swap) or not be connected to the previous phrase (discontiguous). Probabilities are estimated for the three possible orientations: $P(orientation|S,T)$. This probability can be conditioned on both the source and the target, or only on the source, and the orientation can be based on either the previous or the next phrase. These probabilities can be estimated from an aligned corpus using a smoothed maximum likelihood estimation (Koehn, 2009).

Decoding

The task of finding the translation option that maximizes the log-linear model (Equation 2.7) is exponential in the length of the input sentence. Thus heuristic search techniques like best-first search or stack decoding are normally used to estimate the best translation. The main idea is to use a priority queue, where partial hypotheses are stored together with their scores, and where the current best hypothesis is expanded at each step. This priority queue can be pruned to a specific size to reduce time and memory complexity at the cost of risking removing partial hypotheses that would be useful in the end.

One example of a search algorithm used for PBSMT is beam search, which is used in the Moses decoder (Koehn et al., 2007). In this algorithm the target sentence is built from left to right, by expanding any source word phrase. The translation hypotheses are stored in beams, where each beam covers a particular number of source words. Each beam can be pruned independently, based on either histogram pruning, where a limit is set on the maximum number of hypothesis in each beam, or by threshold pruning, where hypotheses are cut based on how much lower score than the best

the|the|DET|DEF boy|boy|N|SING plays|play|V|3-PRES .|.|PUNC|–
pojken|pojke|N|SG-DEF-UTR leker|leka|V|PRES .|.|PUNC|–

Figure 2.4: An example of an English and Swedish sentence represented with factors for surface form, lemma, POS-tags and morphology.

hypothesis in the beam they have. The hypotheses are scored based on their feature function values for the expanded part, and an estimate of the future cost of expanding the hypothesis fully, based on the translation cost and a simplified language model cost (Koehn, 2009).

Parameter optimization

The weights, λ_m, of the log-linear model (Equation 2.7) should reflect the importance given to each of the models. The weights can be optimized on an evaluation metric against a development corpus (see Section 2.3, for a description of some common MT metrics). This process is often called tuning.

A commonly used procedure for performing such optimization is minimum error-rate training (MERT, Och, 2003). It works by translating a set of sentences using some initial weights, producing an n-best list of translation hypotheses. The feature weights are then recalculated, to produce a good ordering of the n-best list with respect to the translation metric scores. The translation step is then repeated with the new weights. These steps are iterated until no new translation hypotheses are found in the translation step or until the improvement is below a certain threshold. Normally MERT is used to optimize the Bleu metric (see Section 2.3).

An alternative procedure was presented by Simard et al. (2005), based on an alternative optimization technique described in Och (2003). By using a smoothed version of the NIST metric, Simard et al. (2005) could use gradient-based optimization, which is not the case for MERT.

2.1.3 Factored SMT

In the models discussed so far, each token in the source text is represented by its surface form. In a factored model (Koehn and Hoang, 2007) each token is represented as a vector of features, which can include linguistically motivated features such as lemmas, POS-tags and morphology, as illustrated in Figure 2.4.

In factored PBSMT an additional type of model, a generation model, can be used. The generation model is only used on the target side, to generate a

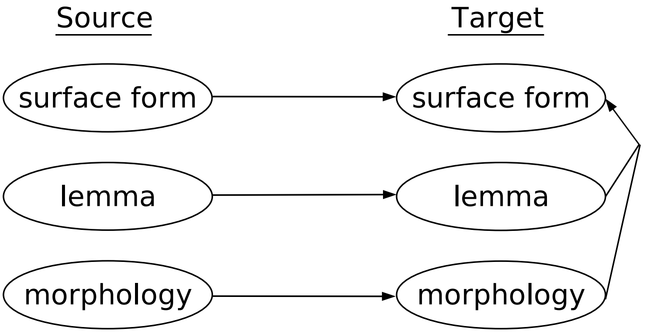

Figure 2.5: Example setup for factored translation

factor from one or more other factors, for instance to generate surface form from lemma and morphology. It can be trained on mono-lingual data. The full translation process is decomposed into one or several translation steps and zero, one, or several generation steps, which is called a decoding path. Factors can also be used in lexicalized distortion models.

Another feature of the factored translation framework is that it is possible to have multiple alternative decoding paths (Birch et al., 2007). This makes it possible to combine a standard translation model from surface form to surface form, with more complex models including generation steps. Figure 2.5 shows an example of such a setup for factored translation, where there are two decoding paths, from surface form to surface form, and a more complex path with two translation models and one generation model.

Factored translation has been used for a number of language pairs in order to target several problems with standard PBSMT. One way to use factors is to have several factors in the target language, and use other sequence models in addition to the ordinary language model. This can improve word order and agreement. Improvements have been demonstrated by using morphologically enriched POS-tags as an extra output factor for translation into German (Koehn et al., 2008; Stymne et al., 2008), and by using supertags for translation from Dutch to English (Birch et al., 2007). Shen et al. (2006) used factors for truecasing, with a target-side generation step for case. Avramidis and Koehn (2008) use source side factors to model case in translation between English and Greek. Morphological models where morphological tags were used both in sequence models and for generation has been used for translation to Czech (Bojar, 2007) and Arabic (Badr et al., 2008). An elaborate model for English to Hindi translation was presented by Ramanathan et al. (2009), where lemmas, suffixes, and semantic relations are used on the source side, and a generation model is used on the target side to combine lemmas and suffixes or case markers to surface form. The setup with several decoding paths can also be used for domain adaptation by combining translation models trained on in-domain and out-of-domain corpora (Koehn and Schroeder, 2007).

2.1.4 Tree-based SMT

The SMT methods discussed so far are flat models that treat sentences as sequences, and do not impose any structure, which can be contrasted to tree-based methods. There has been much research on different types of tree-based models. Such models can either be formally syntactic or linguistically syntactic. Linguistically syntactic models are informed by linguistic theory or annotations, whereas formally syntactic models are hierarchical, but are extracted from raw text. One advantage of these types of models is that they can capture syntactic regularities and reorderings that go beyond the power of PBSMT. A drawback of linguistically syntactic models is that they require tools or treebanks that are not available for all languages.

Formally syntactic models can be induced from plain parallel text, and do not require any parsers or treebanks. Chiang (2005) suggested a hierarchical model based on formal syntax, where a synchronous context free grammar (SCFG) was induced automatically from plain parallel data. Decoding is performed by chart parsing. Other options are the use of head transducers (Alshawi et al., 1997), bracketing transduction grammars (Wu, 1997), or linear transduction grammars (Saers, 2011).

Linguistically syntactic models are based on syntax trees on the target side, string-to-tree models; on the source side, tree-to-string models; or on both sides, tree-to-tree models. String-to-tree models are often based on SCFG (Yamada and Knight, 2002), but have also been based on synchronous tree substitution grammar (Galley et al., 2006), tree-adjoining grammar (Carreras and Collins, 2009), and dependency grammar (Shen et al., 2010). Zollmann and Venugopal (2006) proposed a string-to-tree model similar to that of Chiang (2005), but one that used syntactic categories on non-terminals. In tree-to-string translation, the source side is parsed, in order to guide the translation into target side strings. Tree-to-string models have for instance been proposed based on dependency grammar (Quirk et al., 2005), tree transducers (Graehl and Knight, 2004), and constituency grammar (Huang et al., 2006; Liu et al., 2006). Tree-to-tree models are more powerful with trees on both sides, but they are also computationally even more complex. Again, several different formalisms have been used such as synchronous tree-substitution grammar (Zhang et al., 2007a), synchronous tree-adjoining grammar (Shieber, 2007), and quasi-synchronous grammar (Smith and Eisner, 2006).

While several of these approaches have shown significant improvements over phrase-based models, their search procedures are more complex and it is more difficult to apply a language model efficiently. Due to this some methods do not scale well to large training corpora. Improving efficiency of tree-based models is an active research area, however, for instance there is work on faster language model integration by forest rescoring and cube pruning (Huang and Chiang, 2007).

2.2 Pre- and postprocessing

Pre- and postprocessing techniques are extensions to SMT, which add additional processing modules to the main strategies described in the previous sections. In this section I discuss pre- and postprocessing in general. Strategies used for text harmonization with respect to the focus areas of this thesis are described in more detail in Chapter 3.

2.2.1 Preprocessing

In almost all SMT systems, some pre- and postprocessing is performed. Typically the training data and translation input are tokenized and lowercased. In the tokenization step for languages like English words are separated from punctuation marks and other tokens. For languages like Chinese, which are written without spaces between words, this becomes more difficult, and some type of word segmentation is needed. Lowercasing the data is common in order to reduce sparsity. An alternative is to truecase the data by only changing case of the first word in each sentence (Koehn et al., 2008).

Preprocessing of the bilingual corpora and of the translation input has also been used to target language specific phenomena. Nießen and Ney (2000) described work where they performed a number of transformations on the German source side for translation into English. One of the transformations was to join separated verb prefixes, such as *fahre . . . los/losfahren* (to leave) to the verb, since these constructions are usually translated with a single verb in English. They also split compounds, merged some multi-word phrases, annotated ambiguous function words with POS-tags and replaced some OOV words. Diab et al. (2007) discussed how different types of diacritization of the data influenced translation from Arabic.

It is common to split morphs in a preprocessing step to harmonize morphologically rich languages to languages like English with little morphology. One way to achieve this is by re-segmenting the input by splitting off morphemes. This type of strategy has been successful for translation from Spanish and Catalan (Popović and Ney, 2004), Czech (Goldwater and McClosky, 2005), Arabic (Lee, 2004; El Isbihani et al., 2006; Habash and Sadat, 2006), Finnish (Zwarts and Dras, 2008), and Turkish (Bisazza and Federico, 2009). Which type of morphological segmentation that is best varies with the source language, and in most of the papers above several types of segmentations are investigated with varying results. Often, a 1-best segmentation is used as input to the decoder. However, there are also ways to combine different segmentations in the decoder. Dyer et al. (2008) used a lattice with several segmentations as input to the decoder for translation from Chinese and Arabic. de Gispert et al. (2009) combined n-best lists from decoder runs with different segmentations, and combined them using minimum Bayes risk decoding for translation from Finnish and Arabic.

A similar strategy is to remove morphological distinctions irrelevant to the target language, for instance case distinctions on nouns are not used in English. Removal of such distinctions has been based on Markov random fields (Talbot and Osborne, 2006) and POS-based heuristics (El-Kahlout and Yvon, 2010).

Fraser (2009) proposed a preprocessing method for normalizing German spelling, which was changed after a relatively recent writing reform (Institut für Deutsche Sprache, 1998). He classified texts as belonging to either the old or new spelling, based on the spelling of the complementizer *dass / daß* (that). He then mapped variants of words written differently, such as *duerfen / dürfen* (can) to a single class from which the most common option, based on frequency counts from a corpus was chosen.

2.2.2 Postprocessing

If preprocessing is performed on target language texts prior to training, a postprocessing step of the translation output, where it is transformed back to standard target language is needed. For standard procedures like tokenization and lowercasing, the opposite processes of detokenization and recasing (see e.g., Wang et al., 2006) are performed after translation. For linguistically motivated preprocessing, like morphological processing, there has been much less research on the postprocessing step of morphological merging that is needed if translating into a morphologically rich language, than there has been on morphological splitting for translation from these languages.

Virpioja et al. (2007) split words into morphemes, prior to training, for translation between Finnish, Swedish and Danish. They marked all split modifier parts, with a special symbol. In the postprocessing step, every word that was marked with a symbol was merged with the next word. The translation results measured by automatic metrics were worse when splitting and merging was used than without morphological splitting. However, an error analysis of the result showed other advantages, such as a reduction of untranslated words. No analysis of the merging itself took place. This strategy does have the advantage of being able to merge novel word forms, but has a drawback in that it can merge parts into nonwords if the parts are misplaced in the translation output.

Another study of postprocessing of morphs is El-Kahlout and Oflazer (2006), where translation from English into Turkish was explored. Prior to training, morphs were split and the modifier parts of each word were marked with a symbol and affixes were normalized to base form. In the merging phase, surface forms were generated following morphographemic rules. When the parts were just merged, based on symbols, it gave rise to many badly formed words, and translation results were bad. The reason for this was that the parts were translated out of order. To overcome this to some extent, parts were only merged if the resulting word was accepted by a morphological analyser, ignoring other, redundant or wrong, morphemes. This constraint improved translation, but it was still worse than the baseline without morphological processing. Grouping some of the split morphemes prior to training, i.e., having a lower number of total splits, improved the system above the baseline.

Badr et al. (2008) reported results for translation from English to Arabic. Their best method was a combination merging method where forms were picked at random from the corpus for known combination of morphs and words and were generated based on handwritten recombination rules for unknown forms. This approach gave improvements over a baseline. They also used tags where they combined morphological features with POS-tags as factors in a factored translation model, where surface forms were generated from this information, which gave a small improvement on a large corpus, at the cost of high running times. El Kholy and Habash (2010) extended the merging schemes of Badr et al. (2008) by using the conditional probability and an LM score to choose the best merging option when there were several known variants of a word form. They also backed off to hand-written rules for unknown forms. They showed a little effect of this on an MT task, even though this strategy was the best option in an intrinsic evaluation.

Another approach for treating morphology is to generate the correct morphological form in translation output where only lemmas or other reduced forms are generated. Fraser (2009) used a second decoder run for restoring morphological forms without success. Learning-based methods have however been successful; Minkov et al. (2007) used a maximum entropy model to generate Russian morphology, and Fraser et al. (2012) used conditional random field to generate German morphology.

Nießen (2002) mentioned the need for inverse restructuring if preprocessing is performed on the target side of the corpus. She does not go into any details, but mentions the insertion of the auxiliary *do* in English, and the restoration of separated verb prefixes for German.

There are also postprocessing techniques that do not require any preprocessing. In reranking of n-best lists (Och et al., 2004; Shen et al., 2004; Toutanova et al., 2008) no transformations are performed, but a choice is made between the n best translations produced by the decoder, based on more knowledge than is available to the decoder. Postprocessing has also been used to correct different types of errors in the translation output. Some examples are substitution of words that do not fit in the context (Elming, 2006) and correction of grammatical errors by applying a grammar checker (Stymne and Ahrenberg, 2010).

2.3 Machine translation evaluation

Evaluation of translations is difficult, since there is not one correct answer, but many possible translations that can convey the meaning of a source text in an adequate way. Machine translation evaluation can be either human or automatic. In human evaluation translation output is normally judged in some way by human judges, who preferably should be native speakers of the target language. In automatic evaluation the translation hypothesis is generally compared with one or several human reference translations of the same source text.

2.3.1 Human evaluation

One way for humans to evaluate translation output is to judge them on some scales for adequacy and fluency. This, however, has been shown to be hard, with a low inter-annotator agreement, by e.g., (Callison-Burch et al., 2007), who suggested ranking the translations from different systems either on sentence or constituent level instead. Other evaluation schemes that have been proposed are for instance assessment of acceptability (Callison-Burch et al., 2008), usability (Offersgaard et al., 2008), or measurements based on post-editing MT output, such as keystrokes or mouse clicks (Jäppinen and Kulikov, 1991).

A complement to system-wide human evaluation is to perform an error analysis of the translation output, which aims to identify and classify errors. Error analyses can be large scale categorizations of all types of errors that occur. Such a classification is suggested by Vilar et al. (2006), who used it to evaluate Spanish and English translations. The same classification has been used in other studies, e.g., Avramidis and Koehn (2008) for Greek. Error analysis can also target specific phenomena such as compound translation or noun-phrase agreement (Stymne et al., 2008), Korean verbal heads (Li et al., 2009), or case markers (Ramanathan et al., 2009).

There have also been suggestions for human evaluation that do not require access to the source or a human reference. Examples are studies where reading comprehension questions are asked to participants after reading MT output (Fuji, 1999; Jones et al., 2005), and eye tracking studies of people reading MT output (Doherty and O'Brien, 2009; Stymne et al., 2012b).

Human evaluation is very time consuming and humans often have a low agreement with other humans (Callison-Burch et al., 2007, 2008; Stymne and Ahrenberg, 2012). Thus large-scale human evaluation is performed mostly for larger evaluation campaigns, such as the Workshop of Statistical Machine Translation (see e.g., Callison-Burch et al., 2009). Another drawback of human evaluation is that the effort that goes into evaluation is not reusable; if a system is modified, a new human evaluation is needed.

2.3.2 Automatic evaluation

Most of the commonly used automatic metrics work by comparing the translation output to one or more human reference translations, giving some kind of score that quantifies the closeness to the reference(s). There are a huge number of automatic metrics, but I will focus on the four metrics that are used in the papers of this thesis: Bleu, NIST, Meteor, and PER, of which all are based on surface matching of words, except for Meteor where stemming and WordNet can be used as well. Other approaches to automatic metrics include using POS-tags (Popović and Ney, 2009), syntax (Owczarzak et al., 2007) or deeper linguistic representations such as semantic roles and discourse representation structures (Giménez and Márquez, 2008). It is also possible to combine several metrics (Giménez and Márquez, 2008) or to use machine learning techniques (Duh, 2008).

Bleu (BiLingual Evaluation Understudy; Papineni et al., 2002) is a metric that measures the precision of n-gram overlap with one or several reference translations, and in addition takes into account the length of the translation hypothesis. Equation 2.9 shows the formula for Bleu, where N is the order of n-grams that are used, usually 4, p_n is a modified n-gram precision, where each n-gram in the reference can be matched by at most one n-gram from the hypothesis. BP is a brevity penalty, which is used to penalize too short translations. It is based on the length of the hypothesis, c, and the reference length, r. If several references are used, there are alternative ways of calculating the reference length, using the closest, average or shortest reference length. Bleu can only be used to give accurate system wide scores, since the geometric mean formulation means it will be zero if there are no overlapping 4-grams, which is often the case in single sentences.

$$Bleu = BP \cdot exp\left(\sum_{n=1}^{N} \frac{1}{n} \log p_n\right) \tag{2.9}$$
$$BP = \begin{cases} 1 & \text{if} \quad c > r \\ e^{(1-r/c)} & \text{if} \quad c \leq r \end{cases}$$

Bleu was the first automatic evaluation metric that was shown to correlate well with human judgments. It has become a de-facto standard for machine translation evaluation, even though later studies have shown that other metrics often have a higher correlation to human judgments (e.g., Callison-Burch et al., 2008).

NIST (Doddington, 2002) was developed to target some of the flaws in Bleu. It is also based on n-gram precision and a brevity penalty. However, it does not give equal weight to all n-grams, but less frequent n-grams, which should be more informative, have a higher weight. The brevity penalty is also different. The formula for NIST is shown in Equation 2.10, where $C(w_i \dots w_n)$ is the count of the n-gram $w_i \dots w_n$ in the reference translation(s), L_{sys} is the length of the system output and $\bar{L}_{ref}$ is the average length of the references, β is a constant that is set to make the brevity penalty 0.5 when the word ratio between the system output and the reference is 2/3, and the order of n-grams, N, is normally set to 5.

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum\limits_{\substack{\text{all } w_1 \dots w_n \\ \text{that co-occur}}} \text{Info}(w_1 \dots w_n)}{\sum\limits_{\substack{\text{all } w_1 \dots w_n \\ \text{in sysoutput}}} (1)} \right\} \cdot BP \tag{2.10}$$
$$BP = exp\left\{\beta \log^2 \left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1\right)\right]\right\}$$
$$\text{Info}(w_1 \dots w_n) = \log_2\left(\frac{C(w_1 \dots w_{n-1})}{C(w_1 \dots w_n)}\right)$$

Meteor (Metric for Evaluation of Translation with Explicit ORdering; Banerjee and Lavie, 2005) is different from the above metrics in that it includes recall, not only precision, and only considers unigrams. Fluency is captured by a penalty based on the number of contiguous chunks formed by the matched words. The matching of words is flexible where the matching is performed in stages, starting with surface form and allowing additional matching steps for stems, and for WordNet synonyms. Equation 2.11 shows the formula for Meteor, where P is unigram precision and R is unigram recall based on several matching stages, and α, β, γ are weights. In the original version the weights were instantiated as $\alpha = 0.9, \beta = 3, \gamma = 0.5$. In subsequent versions of Meteor these weights have been optimized against human judgments, both on adequacy and fluency (Lavie and Agarwal, 2007) and on ranking of systems (Agarwal and Lavie, 2008). The original Meteor version can be used for any target language using only surface form matching, but WordNet is only available for English, and the stemmer works only for a restricted number of languages. The optimized versions of Meteor are trained for English, German, French and Spanish.

$$\begin{aligned} \text{Meteor} &= F_{\text{mean}} \cdot (1 - \text{Penalty}) \\ F_{\text{mean}} &= \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \\ \text{Penalty} &= \gamma \cdot \left(\frac{\#\text{chunks}}{\#\text{unigrams\_matched}} \right)^{\beta} \end{aligned} \tag{2.11}$$

PER (Position independent word Error Rate; Tillmann et al., 1997) is one of many different error rates, which are used to calculate the distance of a translation suggestion to a reference translation. The matching is based on the Levenshtein distance (Levenshtein, 1966), the number of insertions, deletions and substitutions needed to transform the hypothesis into the reference. WER (Word Error Rate), is the Levenshtein distance normalized by the reference length. PER is similar to WER, but does not take word order into account. This amounts to comparing the two sentences as bags of words, computing the difference between them, and normalizing by the reference length. One formulation of PER is shown in Equation 2.12 where T_t is the system translation and T_r is the reference sentence (Vogel et al., 2000). Since PER is an error-rate, a lower score is better, and 0 means an identical translation to the reference except for word order.

$$PER = \frac{max(|T_t|, |T_r|) - |T_t \cap T_r|}{|T_r|} \tag{2.12}$$

The main advantage of automatic metrics is that they are cheap and fast to apply, which allows quick testing during system development. They are, however, less informative than human analysis and it is often hard to see exactly what a gain in a metric actually means. Most automatic metrics, including Bleu, are unfair when comparing systems that use different MT architectures, tending to bias in favor of SMT. They are, however, considered useful for incremental development of the same system (Callison-Burch et al., 2006b). In each paper of this thesis I use several metrics, to try to give a broader picture of possible improvements, since the different metrics to some extent measure different aspects of translation quality.

```
Set $c = 0$
Compute actual statistic of score differences $|S_X - S_Y|$
For random shuffles $r = 0, \ldots, R$
    For sentences in testset
        Shuffle variable tuples between system X and Y with probability 0.5
    Compute pseudo-statistic $|S_{X_r} - S_{Y_r}|$
    if $|S_X - S_Y| \geq |S_{X_r} - S_{Y_r}|$
        $c++$
$p = \frac{c+1}{R+1}$
Reject null hypothesis if $p \leq$ specified rejection level
```

Figure 2.6: Approximate randomization test (Riezler and Maxwell, 2005)

Significance testing

It is not straightforward to test the significance of machine translation metrics. The Bleu metric is not decomposable to sentence level, since 4-gram precision is often zero. But even for other metrics that at least in principle work on sentence level, sentence level scores are highly unreliable, and have low agreement with human judgments (see e.g., Callison-Burch et al., 2011, for a sample of MT metrics). Thus it is common in MT to use data-driven nonparametric significance tests (Noreen, 1998) such as bootstrap resampling (Zhang et al., 2004; Koehn, 2004) or approximate randomization (Riezler and Maxwell, 2005). These tests have fewer assumptions than standard analytical significance tests.

In this thesis I use approximate randomization, suggested by Riezler and Maxwell (2005) as being more conservative than bootstrap resampling. It does not require that the samples are representative of the population, which the bootstrap method does. Figure 2.6 shows pseudo-code for performing the approximate randomization to evaluate the difference between two systems. The null hypothesis is that there is no difference between the systems, and the calculation is based on randomly shuffling sentences between the two systems, which should not matter if there is no difference. The test is stratified, by only shuffling translations of the same sentence.

3 Text harmonization: Focus areas and strategies

In this section I describe previous attempts at text harmonization for the four focus areas of the thesis: compounding, definiteness, word order, and unknown words. For each area I first give a brief linguistic introduction to the area in general, especially for the languages treated in this thesis, before presenting previous work relating to SMT. The term text harmonization has not been used in this context, but I consider all types of work that aim to make texts more similar in order to improve SMT to belong to this umbrella term. The description of the focus areas are by no means full introductions to any of these areas. They are intended as a brief background to the text harmonization strategies described in this chapter, and to the thesis work.

3.1 Compounding

Compounding is a type of word formation, where new words are formed by the combination of two or more other words. The majority of compounds are made up of two free morphemes, or stems, that can function as independent words, as in (2). There are, however, instances of compounds where one compound part does not work outside of the compound, as *cran* in *cranberry*. Compounding can be contrasted to other word formation options, such as derivation, where new words are formed from one stem to which affixes are added, as *unsuccessful*, which is formed from the stem *success*, with the addition of the prefix *un-* and the suffix *-ful*.

(2) a. adventure story
    b. competition-friendly
    c. sailboat

As illustrated in (2) compounds can be written orthographically in different ways, as separate words, hyphenated, or as single words. Compounds that are written as single words are called closed compounds, and compounds written as separate words are called open or solid compounds. Languages that has a tendency to mainly use closed compounds are often called *compounding languages*. In many languages compounding is highly productive, and new compounds can be formed readily.

Compounds are sometimes divided into two types: determinative and copulative (Thorell, 1981). In copulative compounds the parts are coordinated and all parts have the same importance, as in the Swedish *blågul* (blue and yellow). These are sometimes further subdivided as copulative and appositional (Bauer, 1983). A copulative compound in this sense, like *player-coach*, which is both a player and a coach, is a hyponym of both parts. In appositional compounds, like *Alsace-Lorraine*, the two parts combine to form the entity referred to by the compound.

In determinative compounds one part, normally the last part in Germanic languages, is the syntactic head of the compound, and the other parts are some kind of modifiers of the head, for instance, a *parliament building* is a building used by a parliament. Determinative compounds can be further subdivided into endocentric and exocentric compounds, depending on the semantic head (Bauer, 1983). In endocentric compounds the syntactic head is the same as the semantic head, which means that the compound is a hyponym of the head, as *parliament building*, which is a type of building. In exocentric compounds the semantic head is not expressed. An example is *paperback*, which is a book with a paper binding.

Compounds can also be classified on the two scales of opacity and lexicalization. Opacity is the degree to which a compound can be understood based on its parts. The meaning of an opaque compound cannot be derived from its parts, as the German *Schnee+Besen* (snow+broom; egg whisk). It can be opposed to transparent or compositional compounds, from which the meaning can be readily derived based on the parts, as the English *airmail*. There is no clear line between transparent and opaque compounds. As noted by Libben et al. (2003) the head and modifiers can have different values for opacity: both parts can be transparent (3a), one part can be opaque and the other part transparent (3b,3c), or both parts can be opaque (3d). Some limited opacity is also shown in words like *blueberry*, which does refer to a blue berry, but to a specific species, not to an arbitrary berry that is blue.

(3) a. car-wash
 b. strawberry
 c. jailbird
 d. hogwash

Lexicalization is related to inclusion of a word in the general vocabulary, or in lexicons, which tend to happen either for very common compounds, or for compounds that have a meaning that is more or less opaque (Dressler, 2006). Another view on lexicalization is that it happens when the productive process of forming the word is not active anymore, as for the derived word *warmth*, where the suffix *-th* is not active in modern English, but the meaning still is clear (Bauer, 1983). On the opposite side of the scale of lexicalized compounds is occasional compounds, which can be formed productively (Hedlund, 2002). Occasional compounds are generally transparent, whereas lexicalized compounds usually are opaque to some degree.

The semantics of compounds can be quite varying and there have been many suggestions of classification schemes for compound semantics. One example is Ó Séaghdha (2007) who suggested six categories, exemplified in (4). The assignment of these categories is not straightforward even for humans; Ó Séaghdha (2007) reported an inter-annotator agreement of 66%, with a Kappa score of 0.62. There have also been attempts at automatically learning the semantics of compounds from corpora (Ó Séaghdha and Copestake, 2009).

(4) a. BE: plastic box (a box made of plastic)
b. HAVE: polio sufferer (a sufferer who has polio)
c. IN: air disaster (a disaster in the air)
d. ACTOR: taxi driver (a driver who drives a taxi)
e. INST(rument): rice cooker (a cooker that is used for boiling rice)
f. ABOUT: tax law (a law about taxes)

Multi-part compounds are normally analyzed as having a hierarchical structure, which influences the semantics. For instance a *small appliance industry* can be analyzed as (small (appliance industry)), whereas *small-appliance industry* can be analyzed as ((small-appliance) industry), giving different semantics, the first being a small industry producing appliances, whereas the second is an industry, of any size, producing small appliances.

3.1.1 Compounds in English

Compounds are common and productive in English. They can have many different parts of speech, including nouns, adjectives, verbs and adverbs (Bauer, 1983), exemplified in (5). Noun compounds are the most common compound type. There are even some closed class compounds, like the preposition *into* or the conjunction *whenever*, but these are not productive. Most English compounds, like those in (5), are endocentric, and have the head as the rightmost part.

(5) a. natural language processing

b. childproof
c. freeze-dry
d. double-quick

As exemplified in (5), English compounds have different orthographic conventions, being either open, closed, or hyphenated. The choice of orthographical realization is inconsistent (Bauer, 1983; Quirk et al., 1985). The pattern can vary even for the same word, which can have several competing realizations (*data set, data-set, dataset*). Quirk et al. (1985) described the general pattern as a progression where newly formed compounds tend to be open, and the spelling moves towards the closed form as the compound becomes established and lexicalized. Bauer (1998) disputes this diachronic view to some extent, claiming that there is no clear correlation between orthography and compound frequency and age. She lists the neologism *airside* and the old compound *college degree* as counter-examples. She mentions other regularities, such as the tendency of multi-part compounds to be open. There is also some variation between varieties of English, with hyphenated compound being more common in British than American English, which prefer open variants (Quirk et al., 1985). Overall, though, closed compounds are mainly lexicalized two-part compounds, while open compounds is the norm for new compounds. English is thus not considered to be a compounding language.

3.1.2 Compounds in German, Swedish, and Danish

Like in English, compounds are both productive and common in German, Swedish, and Danish, and compounds can have varying parts-of-speech. They are also most commonly endocentric, with the head as the last word. Some Swedish examples of compounds are shown in (6). In contrast to English, compounds are generally closed, even though they are sometimes hyphenated, usually when involving a proper name or abbreviation, as in (6e).

(6) a. myggmedel (*mosquito repellent*)
mygga+medel (*mosquito agent*)

b. arbetsskyddslagstiftning (*health and safety legislation*)
arbete+skydd+lag+stiftning (*work protection law creation*)

c. tättbefolkad (*densely populated*)
tät+befolkad (*dense populated*)

d. hamn- och lotsavgifter (*port and pilotage dues*)
hamn- och lots+avgifter (*port- and pilot fees*)

e. Tobin-skatt (*Tobin tax*)
Tobin-skatt (*Tobin tax*)

f. klargöra (*clarify*)
klar+göra (*clear make*)

Compound modifiers can have a different morphological form than the base form of the part as a stand-alone word. The head of the compound, on the other hand, can occur in any paradigmatic form, and does not show any changes specific to compounds. The special modifier forms can differ from the base form in that letters are added and/or removed from it. This change has been called compounding form, connecting element (De: *Fugenelement*, Sv:*fogeelement*), linking suffix, linker, filler, and juncture morpheme. These changes could be additions, such as adding the letter -*s* to *skydd* in (6b), deletions, such as removing the letter -*a* from *mygga* in (6a), and combinations, such as the swap of the letters *e/s* for *arbete* in (6b). There is also a null operation, with no change, such as *lots* in (6d). This type of change in compound form is rare in English, but relatively common cross-linguistically.

Compound forms can coincide with paradigmatic forms, such as genitive and plural. An example of a form that coincides with genitive is *skydds* in (6b), and an example of a form coinciding with plural is the German *Stadien* in (7). An alternative analysis would be to analyze these forms as paradigmatic forms rather than as compound forms. Langer (1998) argues against this, since for German nouns, plural and singular forms in compound modifiers do not always correspond to plural and singular semantics.

(7) Stadienexperte (*stadium expert*)
Stadion+Experte (*stadium expert*)

Some compound parts have different forms in different compounds, exemplified in (8) by the German word *Kind* (child). Most compound parts tend to have only one or very few possible compound forms, with the null operations being the most common. Which compound form a part should have in a particular compound is very hard to predict. There are no rules, but many tendencies, which means that it is hard to formalize them in an automatic system.

(8) `0` Kindphase (*child-caring period*)
Kind+Phase (*child phase*)

`+s` Kindslage (*fetal position*)
Kind+Lage (*child position*)

`+es` Kindesschutz (*child protection*)
Kind+Schutz (*child protection*)

`+er` Kinderfilm (*children's film*)
Kind+Film (*child film*)

+- Ein-Kind-Politik (*one-child policy*)
   Ein-Kind-Politik (*one-child policy*)

Goldsmith and Reutter (1998) mentioned several factors that influence the choice of compound suffixes for German: gender, word-length, phonology, diachrony, and dialectal variety. Kürschner (2003) groups factors that influence choice of compound form for German and Danish into the main categories: semantics, inflection, etymology, derivational patterns and phonology. For Swedish, Thorell (1981) used categories based on declension type. However, even within these categories there are no strict rules, but mainly tendencies of patterns based on factors such as phonology, intelligibility, stylistic level and dialectal influences. The compound suffix also varies with the number of parts in a compound, for instance, the middle part in a left-branching ternary compound is more likely to have an *s*-addition than the same part in a binary compound for Swedish.

3.1.3 Compound processing

Compound treatment has been addressed for translation between German and English by several authors. The most common architecture for translation from German is to split compounds in a preprocessing step prior to training and translation, using some automatic method, for instance in Nießen and Ney (2000); Koehn and Knight (2003); Popović et al. (2006); Holmqvist et al. (2007); Fritzinger and Fraser (2010); Macherey et al. (2011). In this strategy an automatic compound splitter is applied to the training data and to the translation input. In the studies cited above, only one splitting option was given as input to the decoder, which can be problematic in case the splitting is wrong, or if any of the parts are unknown. In Dyer (2009) several splitting options were given to the decoder in the form of a lattice. They could not use lattices during training, and in order to solve this, they doubled the training corpus keeping one part without splits and in the other part they used the best splitting option for each word. Experiments showed that this method is successful for translation from German and Hungarian into English.

For translation into German, Popović et al. (2006) investigated three methods. They split compounds during training and after translation merged compound parts back into full compounds, merged English compounds prior to training instead of splitting German compounds, and used compound splitting only to improve word alignment. Overall the split-merge strategy was the best option, and the merging of English compounds the worst, even though it was still better than a baseline without compound processing.

Compounds have also been targeted in other types of translation systems, such as rule-based systems (Rackow et al., 1992; Gawrońska et al., 1994) and example-based translation (Brown, 2002). There are also suggestions for compound translation in isolation (Tanaka and Baldwin, 2003; Moa, 2005; Garera and Yarowsky, 2008; Bungum and Oepen, 2009).

Compound splitting

Compound splitting is the task of splitting compounds into their component parts. It has also been called decompounding. Alfonseca et al. (2008b) summarized the main strategy generally used for compound splitting in the following steps:

1. For each word, split it in every possible way
2. Calculate a score for each possible splitting option using some weighting function
3. Choose the highest scoring splitting option, which could mean choosing not to split at all, if that has the highest score, or if there are no other valid splitting options

The first step is often performed using some kind of word list, and allowing all splitting options where all of the parts are known words. The word list could either be a dictionary, or it could be compiled from a corpus, which tends to give better coverage, especially for specific domains. It is also possible to use special word lists of known compound parts (Sjöbergh and Kann, 2004). In addition to word lists, special attention needs to be given to compound forms, changes to the form of compound parts, and spelling changes. It is hard to predict where these forms will appear, so a common strategy is to allow them on all modifier parts (Koehn and Knight, 2003). It is possible to constrain the set of splitting options further by imposing different types of constraints, such as limiting the minimum length of compound parts or using POS-based constraints (Koehn and Knight, 2003).

There have been many suggestions of how to rank and score the candidate splitting options. For German, Schiller (2005) used a weighted finite state transducer to choose the most likely split based on probabilities of parts being compound modifiers, and preferring a small number of splits. Holz and Biemann (2008) filtered splitting options based on corpus frequencies and the length of parts. Larson et al. (2000) used a corpus, to calculate how many words that share possible prefixes and suffixes, and split at points where both the suffix and prefix are common. Koehn and Knight (2003) used the highest geometric mean of corpus frequencies of the parts, and a method guided by word alignments with English for choosing the best splitting option.

For Swedish, Brodda (1979) used a rule-based method, based on the observation that consonant combinations at splitting points, such as the sequence *lkk* in (9), are often not found in non-compounds. Another approach based on consonant clusters is described in Kokkinakis (2001). Sjöbergh and Kann (2004) tried a number of features for scoring, including semantic context, component corpus frequencies, syntactic context, POS-tags, and character *n*-grams. Their most successful system combined character *n*-grams with POS-tags and a couple of ad hoc rules.

(9) mjölkko *(dairy cow)*
mjölk+ko *(milk cow)*

Another strategy is to use supervised machine learning to train a classifier based on a corpus annotated with compounds. Alfonseca et al. (2008a) trained an SVM classifier, with features including corpus frequencies, mutual information, and anchor point statistics from web pages. They showed that this method work for several languages, and that training data must not necessarily come from the same language as the test data. Friberg (2007) used memory-based learning, and features based on character *n*-grams. Dyer (2010) used conditional random fields to learn lattice segmentations of compounds. Macherey et al. (2011) presented a language independent unsupervised method, which also automatically learns compounding forms.

Compound merging

Compound merging is the task of combining split compound parts into full compounds. It has to be performed in a postprocessing step if compounds are split in the training data, when translating into a compounding language. The task has also been called recompounding or compound recombination. Merging of previously split compounds for machine translation is much less explored than compound splitting, partly since translation into English is much more common than translation into a language with closed compounds.

Popović et al. (2006) merged compound parts in a postprocessing step after translation into German. The split parts were not normalized, and did not have any type of markup. They used a method based on word lists. Two lists were extracted from the original German training corpus, one of compound parts, and one of full compounds. For every word in the generated output, they checked if it was a possible compound part, and if it was, it was merged with the next word if it resulted in a compound. There is no evaluation of the merging as a separate process, but using it in combination with splitting resulted in improved translation results. Some limitations of the method are that it cannot merge unseen compounds, and that it does not handle coordinated compound parts.

Popović et al. (2006) also tried to merge English compounds prior to training, which they called joining, as an alternative to splitting German compounds. For this they applied two methods:

- POS-based joining: English words corresponding to compounds are usually nouns, therefore each consecutive sequence of English nouns was merged into one word.
- Alignment-based joining: Several English words aligned to one German word were considered possible compound parts, and were merged into one word.

Both these methods resulted in an improvement over a baseline without compound processing, but were worse than using splitting and merging of German compounds.

Fraser (2009) merged split German compounds after translation from English, by applying a second PBSMT system trained on German with split compounds and normal German. Again, this method cannot merge novel compounds. The compound merging component was not evaluated in isolation, but in combination with other morphological processing. The combination had a lower Bleu score than his baseline system. In later work Fraser and colleagues (Fraser et al., 2012) used sequence labeling, based on the method presented in Stymne and Cancedda (2011) and in Paper 1 with positive results.

Koehn et al. (2008) discussed treatment of hyphenated compounds for translation into German by splitting at hyphens and treating the hyphen as a separate token that was merged with the surrounding words after translation. The impact on translation results was small.

Compound merging has also been performed for speech recognition. An example of this is Berton et al. (1996) who extended the word graphs output by a German speech recognizer with possible compounds, by combining edges of words during a lexical search. The best option was then identified from the graph using dynamic programming techniques. Compound merging for speech recognition is a somewhat different problem than for machine translation, however, since the order of parts is not an issue, as compared to PBSMT, where there is no guarantee that the order of the parts in the translation output is correct.

Another somewhat related problem to compound merging, is that of detection of erroneously split compounds in human text, which is faced by grammar checkers. Writing compounds with spaces between parts, as separate words, is a common writing error in Swedish and German. Carlberger et al. (2005) described a system for Swedish that used hand-written rules to identify, among other errors, erroneously split compounds. The rules used POS-tags and morphological features. On a classified gold standard of writing errors they had a recall of 46% and a precision of 39%, for identifying split compounds, indicating that it is a hard problem to find split compounds in free, unmarked text.

3.2 Definiteness

There is no general agreement on a definition of definiteness. Lyons (1999) discussed how definiteness has been defined based on major components of meaning: familiarity, identifiability, uniqueness, and inclusiveness, exemplified in (10–13), with examples adapted from Lyons (1999). He noted that when a definite is used, it should be clear to the hearer which referent is referred to, that is, the hearer should be familiar with the referent. This is exemplified in (10), where we can assume that the hearer is familiar with the bathroom referred to. This is instead sometimes referred to as identifiability, that the use of a definite signals to hearers that they are in a position to identify the referent, as in (11), where it is signaled to the hearer that he can identify the whereabouts of the hammer by looking around the room. Uniqueness is another aspect of definiteness which can apply to singular definites, indicating that they are a unique referent, as in (12), where there is only one possible bride, who might, however, not be familiar to the hearer. For plurals and mass nouns it instead signals inclusiveness, that the totality of the objects or mass in question is referred to, as in (13), where the hearer can assume that all prizes were given out, not just a subset of them.

(10) It is in the upstairs bathroom.

(11) Pass me the hammer, please!

(12) I was to a wedding where the bride wore blue.

(13) She gave out the prizes

Lyons (1999) distinguish between what he calls *simple definites*, which are noun phrases with overtly marked definiteness, for instance by using a definite article, and *complex definites* where definiteness is not marked by a constituent that solely expresses definiteness. Constructions that are generally considered to be complex definites are for instance noun phrases modified by other determiners like demonstratives or by genitives, or proper names.

The realization of definiteness marking in simple definites varies between languages. In fact, in the majority of the world's languages there is no definiteness marking in simple definites (Lyons, 1999). For the large minority of languages that mark them there are several different options, such as free standing forms, like definite articles as in English, affixes that attach to the head, sometimes called bound articles, and phrasal clitics, which are not free-standing words, but can attach to other words than the head noun. It is also possible to mark definiteness on adjectival modifiers, adpositions, or by using different word orders. There are also languages with several possibilities, such as the Scandinavian languages, which have both a free-standing definite article and a definite noun suffix, as well as adjective inflection in definite NPs.

The distribution of the usage of definiteness varies somewhat between languages. Lyons (1999) brings up the examples of generic noun phrases, which tend to be definite in French but not in English, exemplified in (14) with an example from Lyons (1999). Another difference is in reference to body parts in English and Swedish, where English tend to use genitive, whereas Swedish tend to use a simple definite (15).

(14) a. She loves detective novels
b. Elle adore les romans policiers
She loves the novels detective

(15) a. He hit his head on the radiator
b. Han slog huvudet i elementet
He hit head-DEF in radiator-DEF

3.2.1 Definiteness in English and Italian

In simple definites in English and Italian definiteness is expressed by the use of definite articles. English has one article *the*, which is used both in singular and plural (16). There are several definite articles in Italian (17) for different gender, number, and onset of the following word:

- Singular, masculine: il/lo
- Singular, feminine: la
- Plural, masculine: i/gli
- Plural, feminine: le

Several of the articles can also be reduced to *l'* if followed by an article. In Italian the definite articles are also merged with certain prepositions, such as *da* + *il* ⇒ *dal* (from the) or *in* + *gli* ⇒ *negli* (in the).

(16) I bought the book and the pens.

(17) Vedo l'anatra, la capra, i cavalli e le mucche
I see the duck, the goat, the horses and the cows

3.2.2 Definiteness in the Scandinavian languages

Definiteness in the Scandinavian languages can be expressed in two ways: by a preposed definite article, and by a definite suffix of the head noun, exemplified in (18) for Danish.[1] In Danish only one type of definite marking can be used at the same time depending on the context, whereas both types can be used at the same time in the other Scandinavian languages, except Icelandic, giving rise to a phenomenon called double definiteness, or double determination, as in (19) for Swedish. In Icelandic only the definite suffix is normally used in spoken language, and the prefixed article mainly in written and formal language. In the following I will focus on Danish, Swedish, and Norwegian (Bokmål), which are languages treated in the thesis.

(18) Den sjove version af rejsen
The amusing version of trip-DEF

(19) det enorma tempot och de enorma dimensionerna
The enormous tempo-DEF and the enormous dimensions-DEF

The usage of the two types of definite markers varies between noun phrases that are bare, i.e. have no modifiers, and noun phrases with pre-modifiers, for instance adjectives or numerals. In bare NPs, the definite suffix is used in all three languages (20). In pre-modified modified NPs only the definite article is used in Danish, whereas both types of marking are used in Swedish and Norwegian (21). In all languages there is also adjective inflection that is related to definiteness, traditionally referred to as weak and strong endings, as illustrated by the difference in the adjective form in a Swedish indefinite and definite NP (22).

(20) DA bogen
book-DEF

SV boken
book-DEF

NB boken
book-DEF

(21) DA den sjove bog
the funny book

[1] There is no agreement on the status of the latter type of marking, it has been called both a suffix (Hankamer and Mikkelsen, 2002) and a suffixed article (Dahl, 2003). This distinction is not important to the presentation in this thesis, and I will use the term suffix.

SV den roliga boken
the funny book-DEF

NB den morsomme boken
the funny book-DEF

(22) a. ett högt berg
a high-STRONG mountain

b. det höga berget
the high-WEAK mountain-DEF

There are also a few cases where the definiteness marking can vary, for instance in relative clauses, where both types of definiteness marking can be used in all three languages, exemplified for Swedish in (23). The first option, with the article, corresponds to a restrictive interpretation, whereas the second generally would be interpreted as non-restrictive (Hankamer and Mikkelsen, 2002). In Danish, the definite suffix can only be used in NPs that are not pre-modified. In Swedish and Norwegian, however, the definite suffix can also be used in some types of complex definites, exemplified in (24) with demonstratives[2] and genitives. Dahl (2003) identifies three types of words, some relative pronouns like *samma* (same) or *höger/högra* (right), ordinal numerals and superlatives, which are related to a varying use of definiteness marking, with several possible allowed forms.

(23) a. Den sektor som vi talade om är skiftande
The sector that we talked about is varying

b. Sektorn som vi talade om är skiftande
Sector-DEF, which we talked about, is varying

(24) SV Den här boken är bra
This book-DEF is good

NB Han var faren min
He was father-DEF mine

He was my father

The form of the definite article vary with gender and number, in all three languages:

- Singular, common: den
- Singular, neuter: det

[2] There is also a set of Swedish demonstrative that coincides with the definite articles *den/det/de*, which is normally marked by intonation. A noun phrase like *den björnen* is thus grammatical in the interpretation *that bear* but not in the interpretation *the bear*.

- Plural: de

The form of the definite suffix also varies with noun declension and there is some variation, with Swedish examples shown in (25).

(25) a. bulle**n**, tomat**en** och päron**et**
bun-DEF, tomato-DEF and pear-DEF

b. bullar**na**, tomater**na** och päron**en**
buns-DEF, tomatoes-DEF and pears-DEF

The patterns described here are general patterns which are used in written language. There are, however, as Dahl (2003) points out considerable dialectal variants in the usage of definiteness. He identifies a dialect continuum across mainland Scandinavia, where the usage of the definite article is dominant in the south-west, and the use of the definite suffix is dominant in the north-east.

3.2.3 Definiteness processing

There has hardly been any work that has focused on definiteness for SMT. The only work I am aware of is an unpublished report (Samuelsson, 2006), where translation between German and Swedish is investigated. In this work definite articles on the German side were joined with the next word if it started with an uppercase letter. Only unmodified NPs were considered, and the transformations were performed based on surface form; there was no tagging or other analysis. There was no effect of this preprocessing for translation into Swedish. For translation into German, however, there was an improvement on Bleu.

Definiteness is not strictly a morphological issue, even though it can be argued that the Scandinavian definite suffix is a morphological suffix. Even so, the work on morphological processing described in Section 2.2, is somewhat related, especially the work for Arabic, where definiteness is marked by the prefix *Al-* that could be attached to both nouns and other words. For instance, Lee (2004) described a technique where morphs on the Arabic source side were split, and superfluous *Al-* articles were removed based on a set of rules. In translation to Arabic, the definite prefixes have been split into separate words in the training data, and recombined in a postprocessing step (e.g., Badr et al., 2008).

There has been some work on definiteness and articles in connection with rule-based translation. Meya (1990) discussed interlingual representations for definite NPs and Gawrońska-Werngren (1990) and Knight and Chander

(1994) discussed the insertion of articles for translation from Russian and Japanese to English.

3.3 Word order

There is a large variety in the word order of different languages. There are, however, also many features that are shared between many languages. Word order typology is a research area that focuses on finding differences and commonalities between the word order of different languages.

Comrie (1981) brought up a number of parameters that are often used to classify languages for word order, such as constituent order in clauses, the order of adjectives and genitives in noun phrases, and whether a language has prepositions or postpositions. Basic constituent order is a classification of the order of the subject, verb and object. Examples of the three most common constituent orders are shown in (26), with examples from Comrie (1981).

(26) SVO (English) The farmer killed the duckling
SOV (Turkish) Hasan öküz-ü aldı
Hasan ox-ACCUSATIVE bought
Hasan bought the ox
VSO (Welsh) Lladdodd y ddraig y dyn
killed the dragon the man
The dragon killed the man

The order within constituents also varies. In noun phrases, adjectives can either precede the noun as in English (27) or be placed after the noun, as in Italian (28). Genitives can also be placed either in front of the noun, as in Swedish (29), or after the noun, as in French (30).

(27) An interesting book

(28) una lingua difficile
a language difficult
a difficult language

(29) bordets ben
table-DEF's leg
the leg of the table

(30) la patte du chien
the paw of the dog
the paw of the dog

For many languages there are several options for a parameter, and it is hard to classify them into one group, even though there is often a tendency of which order that is the most common. One example is for the order of nouns and genitives in English, where both orders are possible, depending on the type of possessor (31).

(31) a. the child's eye
b. the start of the year

3.3.1 Word order in English and German

English and German are two languages that are relatively closely related. Even so, there are several word order differences between these languages. Maybe the most notable is the difference in the placement of verbs. English is a clear SVO language, with the basic word order subject, verb, object. For German though, the word order is different depending on the clause type (Comrie, 1981). In main clauses, the standard word order is SVO (32a), like in English. If an auxiliary or modal verb is used (32b), the verbs are split, with the auxiliary or modal before the object, and the main verb after the object, leaving the word order classification unclear. In subordinate clauses, the order is SOV (32c), with the verb placed after the object. There are further complications to this scheme though. German is a verb second language, which means that the verb is always placed in the second position in main clauses. This means that in main clauses that have, for instance, an initial adverb, the constituent order becomes VSO (32d). Thus, there are often word order differences between English and German due to the placement of verbs; the difference in placement can be large, exemplified in (33), where both the verb *verlieren* (lose) and other constituents are placed in quite different positions.

(32) a. Das Kind ist froh.
The child is happy.

The child is happy.

b. Er hat Geige gespielt.
He has violin played.

He has played the violin.

c. Ich denke, daß es ein Anfang ist.
I think, that it a beginning is.

I think, that it is a beginning.

d. Jetzt kaufe ich ein Buch.
Now buy I a book.

I buy a book now.

(33) Wir sollten Präsident Pervez Musharrafs Anteil an diesen
We should President Pervez Musharraf's part in these
Geschehnissen nicht aus den Augen **verlieren**.
events not out of the eyes lose.
We should not **lose** sight of President Pervez Musharraf's partial responsibility for this turn of events.

German, as a morphologically rich language, has a freer word order than English. One consequence of this is that it is possible, and quite common, to front most constituents in German in order to emphasize them, which is not always possible in English. For example, it is possible to front objects and prepositional modifiers in German (34), which is quite marginal in English, where often intonation would be used to emphasize the bold parts in (34), which corresponds to the fronted German parts.

(34) a. Seinen Wohnwagen kann man mit der Fähre mitnehmen.
His caravan can one with the ferry take
You can take **your caravan** on the ferry

b. Mit der Fähre kann man seinen Wohnwagen mitnehmen.
With the ferry can one his caravan take
You can take your caravan **on the ferry**

In both languages the placement of adverbs is relatively free with several allowed possibilities, as exemplified in (35). Note that an initial adverb in German, also affects the order of the subject and verb. This variation in itself often leads to shifting positions for adverbs in parallel sentences, as in (36).

(35) EN (quietly) the teacher (quietly) addressed her class (quietly)

DE i. oft gehe ich schwimmen
often go I swim
often I go swimming

ii. ich gehe (oft) schwimmen (oft)
I go (often) swim (often)
I often go swimming

(36) a. **Gestern** war dies erneut der Fall ...
Yesterday was this again the case ...
The same thing happened again **yesterday** ...

b. Als ich **gestern** im Parlament ankam ...
When I yesterday in Parliament arrived
Yesterday when I arrived in Parliament ...

There are also similarities between these two languages. The word order within noun phrases is essentially the same, with adjectives preceding the noun, and relative clauses and prepositional modifiers being placed after the noun. In both languages genitives can be placed both before and after the noun, with placement before the noun being the standard for proper names, and placement after the noun being the standard for inanimate nouns. For nouns referring to humans the placement of genitives is different between the languages, though, with placement before the noun as default for English and placement after the noun as default for German (37).

(37) Johns Sohn, das Buch des Kindes und das Bein des Tisches
John's son, the book of-the child and the leg of-the table
John's son, the child's book, and the leg of the table

Several studies have brought up differences between English and German word order that affects SMT. Nießen and Ney (2004) addressed the word order in questions and the order of separable German verb prefixes. Collins et al. (2005) addressed the order of verbs and subjects, the order of particles, and the negation *nicht*.

3.3.2 Word order in Haitian Creole and English

The word order in Haitian Creole has many similarities to English (Valdmann, 1970). Both languages are SVO languages, with the basic word order subject, verb, object, as illustrated in (38–39). In copulative sentences, the copula is sometimes not explicitly used in Haitian Creole, however (40).

(38) Li pale franse
He speaks French
He speaks French

(39) Mwen se yon etidyan
I am a student
I am a student

(40) Li gran
He big
He is big

(41) festival Baalbek la
festival Baalbek the
the Baalbek festival

(42) nan pwen sila
at point this
at this point

(43) ekipye mwen yo
teammates my PL
my teammates

The main difference between the two languages is the order of elements in noun phrases. In Haitian Creole, the indefinite article is placed before the noun as in English (39), but definite and demonstrative articles are placed after the noun, and the order of other modifiers are also often reversed (41–42). Also with possessive pronouns, the order is generally reversed in the two languages (43).

3.3.3 Reordering

There has been extensive research on how to handle word order differences between languages for SMT. One type of approach is by using stronger translation models than PBSMT, such as different types of tree-based models, see Section 2.1.4, or to use reordering models for PBSMT, as suggested by Tillman (2004). In this section I am concerned with harmonization approaches where one side of the corpus, typically the source side, is transformed to look like the other with respect to word order, sometimes called preordering or pre-translation reordering.

In one type of approach the source side is parsed, and transformation rules are applied to the parse tree. The rules can be either hand-written, which has been explored for German–English (Collins et al., 2005), Japanese–English (Komachi et al., 2006), Chinese–English (Wang et al., 2007), and for several language pairs (Xu et al., 2009). The types of rules created are based on linguistic knowledge of the language pairs, for instance changing the order of some subjects and verbs for translation from German to English. Another option is to learn reordering rules from an aligned corpus (Xia and McCord, 2004; Crego and Mariño, 2007; Habash, 2007; Genzel, 2010). In these papers customized rule extraction strategies were used, but there has also been work using standard machine learning techniques, such as a Maximum Entropy model (Li et al., 2007) or rule-induction learning (Elming, 2008b). These approaches are language independent, and are often effective for several languages, even though results are sometimes only presented for one language pair. Different types of tree structures have been used, such as constituency grammar (Collins et al., 2005), dependency grammar (Xu et al., 2009), and predicate-argument structure (Komachi et al., 2006).

Reordering rules can also be based on POS-tags or chunks. There are hand-written rules based on POS-tags (Popović and Ney, 2006) and automatically learnt rules based on POS-tags (Crego and Mariño, 2006; Rottmann and Vogel, 2007; Niehues and Kolss, 2009) or chunks (Zhang et al., 2007b; Crego and Habash, 2008). Costa-jussà and Fonollosa (2006) described a strategy with two-phase decoding, where the first phase, statistical machine reordering, used SMT methods to change the source word order based on clustered word classes, before a standard decoding step from reordered source language to target. They also showed that clustered word classes worked better than a morphological tagset (Costa-jussà and Fonollosa, 2007). Sometimes several levels of information are mixed, such as POS-tags and chunks (Crego and Habash, 2008), surface form, POS-tags, and phrase structure (Elming, 2008b) or surface form and POS-tags (Rottmann and Vogel, 2007) or parse trees (Xia and McCord, 2004).

Reorderings can be presented to the decoder in different ways such as deterministically choosing the 1-best reordering (e.g., Xia and McCord, 2004) or presenting many possible reorderings to the decoder either as an n-best list (e.g., Li et al., 2007) or as a lattice (e.g., Zhang et al., 2007b). Rules have to be scored in some way to choose the 1-best or n-best option, and sometimes to weigh lattice edges or to prune lattices. Rules have been scored by the relative frequency of rules and the rule length (Rottmann and Vogel, 2007), conditional probability (Habash, 2007), and maximum likelihood estimation based on the performance on training data (Rottmann and Vogel, 2007). Another possibility is to apply rules in a predefined order (Collins et al., 2005), or to choose which rule to apply based on heuristics such as rule length and level of lexicalization (Xia and McCord, 2004). Genzel (2010) created an ordered rule set by iteratively selecting one new rule based on its performance on the training data It is also quite common to use heuristics for pruning rules, such as requiring a rule to occur a minimum number of times in the training data (Niehues and Kolss, 2009). Reordering rules can also be integrated into the decoder (Tillmann, 2008) or scored by the decoder (Elming, 2008b).

Another difference between approaches is to which data the reorderings should be applied. They can either be applied only to the translation input (e.g., Elming, 2008b), or also to the SMT training data (e.g., Popović and Ney, 2006). Rottmann and Vogel (2007) found that it was better to reorder the training data based on the learnt reordering rules, than to use the original order or alignment-based reordering, a reordering created by moving words based on the word alignments, for the training data. Zhang et al. (2007b) had a small improvement by using a combination of alignment-based reordering and original order in the training data, compared to using only the original order. Another option for using reordering in the training data was presented by Niehues et al. (2009), who directly extracted phrase pairs from reordering lattices, and showed a small gain over non-reordered training data.

Most preordering strategies have been investigated for phrase-based or n-gram-based SMT. There are a few studies that has shown that it is useful even for more complex SMT strategies. Xu et al. (2009) found that their preordering strategy worked both for phrase-based and hierarchical SMT and Wang et al. (2010) showed that a syntactic SMT system can gain from preprocessing such as parse-tree modification.

In all the above studies the reordering was performed on the source side. However, Na et al. (2009) presented a study where they reordered the target side. The target side of the training data was reordered based on alignments, and based on this the SMT system and a non-projective dependency parser were trained. In a postprocessing step to monotone decoding, they parsed the translation output, and adjusted it based on a local tree order model.

A different way to use reordering is to use it only to improve the word alignment, and move the words back into original order after the alignment phase. This approach was used by Holmqvist et al. (2009) with alignment-based reorderings and by Carpuat et al. (2010) addressing the order of subjects and verbs in Arabic.

3.4 Unknown words

Unknown words or out-of-vocabulary words (OOVs), are words that are unknown to a natural language processing system. In the context of SMT, they are words that occur in the texts to be translated, but do not occur in the training data of the SMT system, which is of a limited size.

In general, words have very unequal frequencies, with few words having a very high frequency and many words having a very low frequency. Manning and Schütze (1999) notes that in a corpus made up of a novel, the 100 most common word types, that is, unique words, account for over half of the tokens, that is, individual instances of a word, whereas nearly half of the word types occur only once. This type of distribution is typical for text corpora. Zipf's law says that the frequency of a word is inversely proportional to its rank, the position in a list of words ordered by frequency. Zipf's law is generally not regarded as a law, but rather as a characterization of word distribution, and certain other empirical facts (Manning and Schütze, 1999). The long tail of uncommon words means that even if a very large corpus is used to train an NLP system, there will still be words that are unknown to the system, when translating unseen texts.

There are several aspects of a language or a text that can contribute to a high number of OOVs. One such aspect is the morphological complexity of a language. If a language has many morphological affixes, the number of word types will increase. English is not morphologically complex, for instance there are only two morphological forms of nouns, singular and plural. More morphologically complex languages also encode other distinctions on words, such as number, gender, case, and definiteness for nouns. In Swedish, for instance, there are eight forms of nouns, accounting for number, case, and definiteness, with two values for each. This is by no means extreme; there are for instance languages with a high number of cases, like Finnish, with 14–15 cases. In a morphologically complex language there are thus more possible forms for each word stem, making it more likely that some of these forms will appear as unknown words. As shown in Paper 1 (Tables 5–6), German and Swedish have nearly three times more word types than English for a given corpus. It has also been shown that the morphological complexity of a language influences the quality of PBSMT negatively (Birch et al., 2008).

3.4.1 Non-standard orthography

Another issue that can affect the occurrence of OOV is the use of non-standard orthography in a corpus. If there for some reason is variation between the spellings of the same word, it can contribute to a high OOV-rate. In this section I will briefly discuss two issues, the orthography of Haitian Creole, which is less standardized than European languages, and SMS language for which there are different writing norms than for other texts.

Haitian Creole orthography

Haitian Creole is a relatively young language. It is a creole language, a language that developed as a result of the need for communication between groups of people speaking different languages, often caused by European colonization. It started to develop in the 17th and 18th centuries, and its basic grammatical structure is generally considered to have West-African roots, whereas most of the vocabulary is based on French (Schiefflin and Doucet, 1992). While there are a few written texts from as far back as the 17th century, the first standardized orthography did not appear until 1924, and an official orthography was only adopted in 1980, after Haitian Creole was introduced in schools in 1979. French was the only language used for education, and in government, up until then, and Haiti did not officially become French–Haitian Creole bilingual until 1987 (Schiefflin and Doucet, 1992). Since 1924 up until 1980 there were 11 different major spelling systems suggested.

There is some dialectal variation in Haiti, but the main dividing line between language variants is between sociolects. There is a clear dividing line between the language spoken by the elite, a bilingual minority constituting some 7% of the population, and the monolingual masses (Schiefflin and Doucet, 1992). This variation is reflected in the spelling, and in the struggle to find a unified orthography. Schiefflin and Doucet (1992) identified two extreme standpoints when creating orthographies: a phonemic approach, which advocates a simple mapping between sounds and letters, with the goal of ease of learning, and an etymological approach, which takes the French origin into account, and which minimizes the differences in spelling between the two official languages.

This variation in writing systems and sociolects leads to quite a big variation in actual spelling. Allen (1998) defined Haitian Creole as a *vernacular language* that is in the process of normalization and standardization. He noted that there is a lot of variation in the spelling, especially of common words, sometimes even in the same text. He identified three factors that cause most of this variation: alternations in vowel height, e.g., *e* or *è*, alternations between *r* and *w*, and alternations between oral and nasal vowels. For instance he found 16 different spellings of the word meaning *week* in different corpora, including *semèn, semenn, senmenn*, which differ in vowel height and nasalization, and variations such as *presyon, pwesyon* for the word meaning *pressure*. He also identified words where differences in these aspects result in different words rather than spelling variants, such as *bra* (arm) and *bwa* (tree).

Another irregularity in Haitian Creole is that there are two registers for pronouns: the high register that uses full forms, and the low register that uses contracted forms. As an example, the first person singular pronoun (*I/me/mine*) is written as *mwen* in its full form, but in its contracted form as any of *m', 'm, m*, depending on its context (Lewis, 2010). It is also common to use contracted forms for other combinations of words, such as *poum* for *pou mwen* (for me) (Lewis, 2010) or *lavel* for *lave li* (to wash [it]) (Munro, 2011). Schiefflin and Doucet (1992) noted that the usage of hyphens and apostrophes is optional in the standard orthography, and thus very unstable. As noted by several researchers, the large orthographic variation in Haitian Creole is challenging when building NLP systems (e.g., Allen, 1998).

SMS language

SMS is an acronym for *Short Message Service*, and is used for sending short text messages between mobile phones. Using Latin scripts, each text message is limited to 160 characters unless sending messages in multiple parts. On traditional mobile phones, SMSs are typed using a limited number of keys, with 3-4 letters on each key, making it necessary for users either to press keys numerous times, or to use predicative software like T9, which guess which word is intended. On modern smart phones small keyboards, either hardware of software keyboards, can normally be used for typing SMSs.

Several factors, including the limit of SMS length, the cost of SMSs, and the arduous text input methods, contribute to the fact that people tend to express themselves more concisely in SMSs than in other types of written texts. Hård af Segerstad (2002) mentioned several types of techniques that are used to reduce the lengths of text messages, including syntactic reductions, like deletions of function words or subjects, lexical short forms, like abbreviations and acronyms, omissions of punctuation and spaces, and symbols replacing words. She also identified instances of spelling that resembled spoken rather than written forms, and found that it was common to use either only lower-cased or only upper-cased letters in messages, rather than standard casing. Similar phenomena are reported in many other studies. Pennell and Liu (2011) exemplified SMS text, as in (44), where several of the above techniques are used, and also the removal of vowels and the use of numbers in words.

(44) a. Rndm fct bout wife: n the past 10 yrs I can cnt on one hand the num Xs she's 4gotn to unlock my car door

b. OMG I LOVE YOU GUYS. You pwn :) !!!

There has been much work on normalizing non-standard words in NLP, both in general (e.g., Sproat et al., 2001) and of SMS messages (e.g., Pennell and Liu, 2011). SMS normalization has been attempted using PBSMT both on the word-level (Aw et al., 2006) and character-level (Pennell and Liu, 2011), as well as several other techniques, including hidden Markov models (Choudhury et al., 2007), speech-recognition-based techniques (Kobus et al., 2008), and a combination of lexical and phonological edit distance (Han and Baldwin, 2011). Han and Baldwin (2011) also trained a classifier for deciding which words are ill-formed and in need of normalization.

3.4.2 Online OOV-handling

There are two major ways to tackle the issue of OOVs for SMT systems. One way is to explicitly address the OOVs found in the translation input and add translations for it in some way, which is called online OOV-handling by Habash (2008). Another way is by increasing the coverage of a system by using techniques for reducing sparsity in general. In this section the main focus is on online OOV-handling.

Processing of OOVs in the input texts can take different forms. One approach is to replace unknown words in the translation input with known equivalents. The replacement can be either a 1-best replacement (Arora et al., 2008), an input lattice (DeNeefe et al., 2008) that allows several options, or several different 1-best replacements can be used as input to the decoder in separate runs (Mirkin et al., 2009). Another possibility is to identify translations of OOVs and add them to the phrase-table (Langlais and Patry, 2007; Habash, 2008). Entries can be either modifications of existing entries, by replacing a known alternative in a phrase with an OOV, or completely new entries, for instance containing transliterations of OOVs (Habash, 2008). There has also been work on replacing unknown words in the translation output in a postprocessing step (Eck et al., 2008; Paul et al., 2009). Arora et al. (2008) also recognized that using standard PBSMT techniques, it is possible to have words in the phrase-table without entries as single words, due to either missing or inconsistent alignments, and suggested a technique for extending the phrase-table with such entries.

Many different types of techniques have been suggested for finding alternatives to OOVs that are in-vocabulary (INV). Alternatives can be found based on morphological likeness (Yang and Kirchhoff, 2006; Habash, 2008), similar spelling (Habash, 2008; DeNeefe et al., 2008), transliteration (Habash, 2008; Hermjakob et al., 2008; Paul et al., 2009), dictionary lookups (Habash, 2008; Eck et al., 2008), analogical learning (Denoual, 2007; Langlais and Patry, 2007), character-based translation (Vilar et al., 2007; Zhang and Sumita, 2008), word splitting (Yang and Kirchhoff, 2006; DeNeefe et al., 2008), paraphrasing (Callison-Burch et al., 2006a), and WordNet synonyms and hypernyms (Mirkin et al., 2009). These methods have some overlap, for instance, character-based SMT can be viewed as a type of transliteration, and analogical learning addresses morphology to some extent. Habash (2008) noted that there are different types of OOVs, and that different methods are best suited to handle them. He showed that a combination of different methods was better than any one of the individual methods.

There have been some different options suggested for finding morphological alternatives to OOVs. Yang and Kirchhoff (2006) recursively applied stemming and compound splitting to unknown German words. Habash and Metsky (2008) learnt which prefixes and suffixes in Urdu that expressed information irrelevant for English, such as case for nouns, based on words with the same translation in the phrase-table. They then applied changes for these affixes for OOV words. Habash (2008) used a similar technique for Arabic. Arora et al. (2008) stemmed content word OOVs, and then generated all possible forms based on morphological rules for nouns, adjectives, or verbs. As a backup model they matched stemmed forms to stems from the training corpus.

The identification of spelling alternatives to OOVs is usually based on edit distance operations at character level. Edits can be insertions, deletions, substitutions, and sometimes inversions of characters. Habash (2008) used this strategy for Arabic–English translation, allowing maximum one edit operation per word. He also limited the character substitutions that were allowed. He recycled entries in the phrase-table by adding new entries where INV words were replaced by OOVs. The original weights of the phrase-table entries were kept, and no scoring of the edit distance operations was performed. DeNeefe et al. (2008) created lattices with spelling alternatives for both OOVs and singelton words. Besides the standard edit operations they allowed insertions and deletions of spaces in addition to other letters, and removal of characters not in Arabic orthography. The number of edit operations was limited to one per word, and there was no scoring of edits or weighing of lattice arcs. Bertoldi et al. (2010) also addressed translation of misspelled words, but only experimented on artificially corrupted text. They did not target OOVs explicitly, but directly applied a correction model to a full sentence to create a confusion network, allowing any word to be changed. Substitutions were based on character-level edit distance operations, and scored by the distance between characters on a keyboard and a character n-gram model. In most studies only single unweighted edit operations are used. Even though it has been claimed that most spelling errors are due to a single operation (Damerau, 1964), it has been shown that spell checking can be improved by allowing several edit operations (Brill and Moore, 2000), and also by weighting edit operations individually (Church and Gale, 1991).

Hewavitharana et al. (2011) tried to use spelling normalization on the full training corpus for Haitian Creole, in order to generally reduce OOV rates. They allowed one edit operation per word, if the resulting word was found in a French dictionary, since French is historically related to Haitian Creole, and there are much more French data available. They applied this strategy either to all words in the training corpus, or to rare words, but did not beat the baseline without any spelling normalization.

Some attempts have been made to weigh the alternatives found for OOVs. Mirkin et al. (2009) investigated both source side and target side features. On the source side they used the frequency of the alternative words, as did Arora et al. (2008), and they also used several methods for evaluating how well the alternative fitted into the context: source side language model probabilities, latent semantic analysis, and naïve Bayes. They found that context sensitive source side features could be used to filter the number of alternatives without harming the result.

The methods for morphology and compound segmentation described in Sections 2.2.1 and 3.1.3 are also useful for reducing the number of OOVs in the input, since they reduce data sparsity. For instance, Stymne (2008) showed that compound splitting of the German source led to a 50% reduction of OOVs for translation from German and Popović and Ney (2004) reported reduced OOV-rates of different sizes when using morphological processing of Spanish, Catalan, and Serbian. Factored SMT with generation models, such as that used in Ramanathan et al. (2009), can also reduce the number of OOVs since it can generate unknown word forms. For related languages it is possible to reduce OOV rates by the use of character-based decoding (e.g., Tiedemann, 2012).

4 Experiments and results

In this chapter I first give a general overview of the research approach, and of the main work in the thesis. I then describe the tools and resources used in the experiments. This is followed by individual summaries of the seven papers included in the thesis.

4.1 Research approach

The goal of this thesis is to improve PBSMT by using text harmonization strategies, with a focus on compounding, definiteness, reordering, and OOVs. The research is experimental, and for each area I have implemented systems that carry out some type of text harmonization, and evaluated its effect for translation between different languages.

Text harmonization is achieved by a preprocessing step where the training data and/or the translation input are transformed in one or more ways. I have either used heuristics or machine learning techniques for the transformations. In most of the studies the transformation rules are based on shallow linguistic knowledge, mainly POS-tags, but for the studies on reordering I also used parser-based tags.

I have chosen to mainly use shallow linguistics, such as POS-tags since there are POS-taggers available for many languages, and they are generally fast and have a high precision. For some languages, however, such resources are not available, and alternatives are needed, and I discuss clustered word classes as an alternative for handling reordering issues. In a number of studies I make use of factored translation models (Koehn and Hoang, 2007), which allow linguistic information such as parts of speech to be used in the translation process.

In most of the studies I focus on translation from English into a subset of the other Germanic languages. I think that these languages are a relevant sample of languages. They are both more morphologically complex than English to a varying degree, and the word order differ to some extent, with mostly local differences between English and Scandinavian, and also long distance differences with German, especially for verbs. I also investigate translation from Haitian Creole into English, using an SMS corpus, which

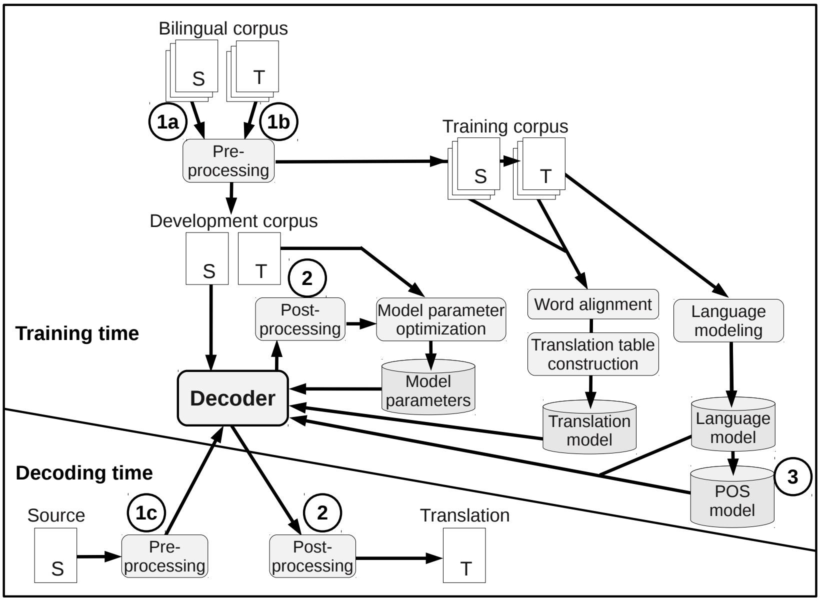

Figure 4.1: Overview of the PBSMT architecture. The areas that I have focused on are marked by circles: (1) preprocessing, (2) postprocessing, and (3) POS-based sequence models.

is an interesting case since it contains non-standardized language in need of harmonization. Haitian Creole is also interesting since it is a less-resourced language, for which there are few, if any, tools and resources such as taggers, parsers, and treebanks.

Figure 4.1 shows an overview of training and decoding with the PBSMT approach, with circles marking the main components I used in my work:

1. Preprocessing
   a) source side of training data
   b) target side of training data
   c) translation input
2. Postprocessing
3. POS-based sequence models

Preprocessing of training and/or test data is the most straightforward way to achieve text harmonization, since one text can easily be transformed to

Table 4.1: Techniques used in Paper 1–7. The focus of the papers are Comp(ounding) in Paper 1, Def(initeness) in Papers 2–3, Reo(rdering) in Papers 4–5, and OOV-handling in Papers 6–7. A capital X indicates that the technique is an important topic of the paper, and a lower-case x that it is used but not as a main topic of the paper.

	Comp	Def		Reo		OOV	
	1	2	3	4	5	6	7
Preprocessing, source		X	X	X	X	x	x
Preprocessing, target	X		x				
Preprocessing, input		X	X	X	X	X	X
Postprocessing	X		x			x	
POS sequence models	X					x	

become similar to another in some respect in this step. When preprocessing is performed on the target side, it becomes necessary to perform a postprocessing step in order to change the processed SMT output. Postprocessing can also be used for text harmonization on its own, for instance by changing the number formatting of the output. POS-based sequence models can be used with a factored decoder, see Section 2.1.3. They are not central to text harmonization, but they can be used in addition to the other techniques.

Table 4.1 shows which techniques are used in the seven papers. The main technique I used is preprocessing, of different kinds. In all papers except Paper 1 the source side of the training data and translation input are modified. In Papers 2–5 these modifications match between training and test, but in Papers 6–7, the input modifications are mostly for OOVs, which are not addressed in the training data. In Paper 1, which addresses compounding on the target side, the target side is preprocessed, making a postprocessing step necessary. This combination of modifications is also investigated to a limited extent for definiteness in Paper 3, but without success. The main use of POS-based sequence models is in Paper 1, where it is used to improve the order of compound parts.

4.2 Tools and corpora

A number of external tools and resources were used in this work. The training and running of the MT system used the Moses toolkit (Koehn et al., 2007) and Matrax (Simard et al., 2005). For language modeling I used the SRILM toolkit (Stolcke, 2002) and KenLM (Heafield, 2011). Word alignments were created using GIZA++ (Och and Ney, 2003). As part of the preprocessing I used the POS-taggers TreeTagger (Schmid, 1994), RFTagger (Schmid and Laws, 2008) Granska tagger (Carlberger and Kann, 1999), and a hidden Markov model tagger (Cutting et al., 1992). I also used a commercial dependency parser (Tapanainen and Järvinen, 1997).

Moses (Koehn et al., 2007) is an open source toolkit for phrase-based SMT that contains a decoder. In addition, Moses contains scripts for creating translation and lexicalized reordering models, and for tuning feature weights. Moses allows factored translation (see Section 2.1.3). It has support for using factors in the translation and distortion models, in additional language models, and in generation steps on the target side. Matrax (Simard et al., 2005) is a phrase-based decoder, which allows discontiguous phrases. Matrax did not allow factored decoding, but I extended it with the possibility to use target side factors. Matrax has an internal language model implementation.

SRILM (Stolcke, 2002) is a toolkit for building and applying language models. The toolkit implements several smoothing methods, including the two methods used in the experiments: modified Kneser-Ney (Chen and Goodman, 1999) and Witten-Bell (*Method C* in Witten and Bell, 1991). KenLM (Heafield, 2011) is an efficient language model inference library that is included in Moses.

GIZA++ (Och and Ney, 2003) is a word-alignment tool that implements IBM model 1–4 (Brown et al., 1993), an HMM-based model that can replace IBM model 2 (Vogel et al., 1996) and parameter smoothing. It produces unidirectional one-to-many alignments between two languages. In the experiments, GIZA++ runs 5 iterations each of model 1 and the HMM model, and 3 iterations each of model 3 and 4. All word alignment is performed on surface forms.

To be able to use POS-tags as a factor and for preprocessing, the training texts have to be tagged. I used TreeTagger (Schmid, 1994) for German and English, RFTagger (Schmid and Laws, 2008) for German, Granska tagger (Carlberger and Kann, 1999) for Swedish, and a tagger based on Cutting et al. (1992) for Danish, Norwegian, and Swedish. All taggers are based on Hidden Markov models, which TreeTagger and RFTagger use in combination with decision trees. The Granska tagger and RFTagger produce morphological analyses, with information such as gender and number for nouns and tense for verbs. The Granska tagger was developed for grammar checking, and makes a few tokenization choices that are not suitable for translation, so the output from it is postprocessed in order to handle time expressions and coordinated compounds. In the reordering experiments I used a commercial dependency parser (Tapanainen and Järvinen, 1997).[1]

[1] Connexor Machinese Syntax, `http://www.connexor.com/nlplib/`

Table 4.2: Corpora, decoder, and language pairs in the papers in the thesis. News is news corpora from WMT workshops, Auto is the automotive corpus, SMS the Ht–En SMS corpus, and Misc variable other corpora for Ht–En.

Paper	Focus	Language pairs	Decoder	Corpora
1	Comp	en⇒{de,sv,da}	Moses, Matrax	Europarl, Auto
2	Def	en⇒da	Matrax	Europarl, Auto
3	Def	en⇒{sv,da,nb}, it⇒da	Moses, Matrax	Europarl, Auto
4	Reo	en⇒de	Moses	Europarl
5	Reo	en⇒de, ht⇒en	Moses	Europarl, SMS
6	OOV	en⇔de	Moses	Europarl, News
7	OOV	ht⇒en	Moses	SMS, Misc

Most experiments were performed on the Europarl corpus (Koehn, 2005), which contains transcriptions of the proceedings of the European Parliament in eleven languages, including English, German, and Swedish. Europarl is sentence aligned using the algorithm by Gale and Church (1993). The current release of Europarl, v6, contains around 1,700,000 sentences for the language pairs used in this thesis, but in order to reduce training times of the PBSMT system, I used smaller partitions of Europarl in many experiments. I also used corpora provided for the Workshops on Statistical Machine Translation (see e.g., Callison-Burch et al., 2011), that included mono- and bilingual news corpora for German–English, and a number of corpora for Haitian Creole–English. The main corpus I used for Haitian Creole is an SMS-corpus of anonymized messages sent after the 2010 earthquake gathered by a consortium of volunteer organizations, *Mission 4636*. I also used a small automotive corpus extracted from a translation memory from automotive manuals, for translation from English to Swedish, Danish, and Norwegian.

Table 4.2 gives an overview of which decoders and corpora that were used in the papers of the thesis. Europarl is used in all papers, except for the experiments on Haitian Creole. Matrax is mainly used as a decoder for the automotive corpus, except in Papers 2 and 3, where it is also used for Europarl. Some samples from the publicly available corpora are shown in Table 4.3.

4.3 Paper summaries

In this section I will summarize the main work and conclusions of each of the seven papers in the thesis.

Table 4.3: Examples from some of the corpora used in experiments in the papers of the thesis

Corpus	Language 1	Language 2
Europarl En–Da	Workers are facing a massive attack on their employment and social rights.	Arbejdstagerne står over for det allerstørste angreb mod deres arbejdsmæssige og sociale rettigheder.
En–De	That is why I believe that we need a different policy mix for a new policy of full employment.	Deshalb glaube ich, daß wir einen anderen policy mix für eine neue Politik der Vollbeschäftigung brauchen.
En–Sv	The amendment to the directive on todays agenda does not therefore affect the existing harmonization of the transport of dangerous goods in the Community.	Den ändring av direktivet som i dag står på föredragningslistan innebär alltså ingen förändring i den standardisering av transport av farligt gods som gemenskapen har i dag.
It–Da	E' assolutamente sproporzionato e non aiuta certo il processo di pace.	Det er helt ude af proportioner og fremmer på ingen måde fredsprocessen.
News En–De	The investment was a mistake, but I learned a lot more than I would have from a success.	Die Investition war ein Fehler, allerdings lernte ich so sehr viel mehr, als wenn ich Erfolg gehabt hätte.
	This year's storms in central and southern China produced the worst winter weather in a half-century.	Die diesjährigen Stürme in Zentral- und Südchina sorgten für den schlimmsten Winter seit einem halben Jahrhundert.
SMS Ht–En	Nom pam se [FIRSTNAME][LASTNAME] mwen se yon chofeur mwen gen 13 zan esperians si nou bezwenm nou ka relem nan numero s a telef.[PHONENUMBER] adrès mwen cl	My name is [FIRSTNAME][LASTNAME]. I am a driver. I have 13 years of experience. Should you need me, I can be reached at [PHONENUMBER] my address
	Si ta gen lapli ki sa pou m fè?	If there is rain what am I to do?

4.3.1 Paper 1

Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2012a. Generation of compound words for statistical machine translation into compounding languages. Submitted manuscript

In this paper we investigated several methods that address compound processing for translation into compounding languages:

- Coalescence: how can the order and adjacency of compound parts be improved for split compounds?
- Compound merging: how can we merge compounds and both produce novel compounds and minimize the risk of erroneous merges?
- Compound splitting: how do differences in splitting strategies influence compound translation?

We performed experiments on two corpora, for translation into German, Swedish, and Danish, all of which have closed compounds. The major goal was to be able to produce novel compounds to overcome the sparsity of compounds in corpora.

To address the issue of coalescence, we used sequence models on customized tagsets. A standard POS-tagset was modified in several ways, the most successful being by adding tags for split compound modifiers, which were based on the compound head, such as noun-modifier for noun compounds. This tagset could then be used as an output factor in a factored translation model. We showed that this strategy led to a large reduction in the number of misplaced compounds, while still producing a higher number of compounds than the baseline system.

For compound merging we developed a new heuristic strategy that was based on matching of tags from the customized tagsets, and improved an existing list-based strategy by several types of constraints. We also showed how sequence labeling could be used to merge compounds, and suggested a useful feature set. These methods performed on par with or better than previous suggestions. In addition, the POS-based heuristic and the sequence labeler can produce novel compounds, which was not the case for previous algorithms.

We also showed that the type of compound splitting used influences the translation results, and that we could not use the results of intrinsic evaluations of compound splitting to predict the success on the machine translation task. This finding is consistent with previous research for the opposite translation direction.

Overall the results were on par with or better than the baseline for several datasets and language pairs. But we also showed that we could increase the number of compounds in the translation output compared to not processing compounds.

4.3.2 Paper 2

Sara Stymne. 2009c. Definite noun phrases in statistical machine translation into Danish. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 4–9. Odense, Denmark

In this paper I discussed how definiteness can be handled for translation from English to Danish. I used a POS-based preprocessing method on the English source side that transformed English definite noun phrases so that they look like Danish. There were two simple transformations:

- Remove definite articles that do not have an equivalent in Danish
- Add a definite suffix to nouns that correspond to Danish suffixed nouns

Both these transformations were applied to bare NPs. Modified NPs have the same structure in Danish and English, and were not modified. For compound nouns, the definite marker was put on the head of the noun, as in (45).

(45) Original: the apple trees in the old garden
Modified: apple trees#def in the old garden
Danish: æbletræerne i den gamle have

I performed experiments on two corpora, Europarl and a small automotive corpus, using the Matrax decoder. I also investigated the interaction between using discontiguous phrases, gaps, in Matrax and definiteness processing. When gaps were used there was a large improvement on both corpora, of 5.4 Bleu points on the automotive corpus and 4.2 Bleu points on Europarl. By not allowing gaps, there was an improvement on the baseline in both cases. With definite processing, there were differences between the two corpora. On Europarl, there was an improvement again, and a similar result to using definite processing with gaps. On the automotive corpus, the result was similar to the baseline without gaps. Overall, though, the best strategy was to use both gaps and definite processing.

This interaction between the types of phrases used and the preprocessing strategy is interesting, and should merit more research. A small error analysis also showed that the preprocessing led to many different types of changes, besides affecting the definiteness construction, such as influencing word choice and word order.

4.3.3 Paper 3

Sara Stymne. 2011a. Definite noun phrases in statistical machine translation into Scandinavian languages. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 289–296. Leuven, Belgium

This paper is an extension of Paper 2 to new language pairs. I investigated how definiteness processing could be used for an additional source language, Italian, and additional target languages, Swedish and Norwegian. The same two operators as above were again used to transform English definite NPs to look similar to the Scandinavian languages. For Italian I also moved adjectives that were placed after the noun, in order to better resemble the Danish structure.

For translation to Danish, the same strategy as in Paper 2 was effective also when combined with compound processing for translation from English. Also for translation from Italian there were significant improvements for translation to Danish.

For translation from English to Swedish and Norwegian, I first applied essentially the same strategy as for translation to Danish, with the exception that definiteness markers are used also in modified noun phrases, to resemble the structure in these languages, as shown in (46) for Swedish. This was, however, not successful neither for Swedish nor for Norwegian. I thus investigated a different strategy where I only removed definite articles, and did not add any markup on definite nouns. This strategy led to improvements on both languages. One explanation for the failure of the initial strategy is that definite markers are used in other types of phrases than pure definite phrases in Swedish and Norwegian, for instance in certain demonstrative phrases, where they are not used in Danish. I also investigated target side processing for English–Swedish translation, but with no positive results.

(46) Original: the small button closes the door
Modified: the small button#def closes door#the
Swedish: den lilla knappen stänger dörren

This study shows that strategies that are developed for one language pair cannot always be applied in the exact same way to other language pairs. With some modification the strategy was, however, successful for all investigated language pairs.

4.3.4 Paper 4

Sara Stymne. 2011b. Iterative reordering and word alignment for statistical MT. In *Proceedings of the 18th Nordic Conference on Computational Linguistics (NODALIDA'11)*, pages 315–318. Riga, Latvia

In this paper I investigated the effects of iterating the learning of reordering rules and word alignment. The intuition behind the iterative approach is that both these processes could potentially benefit from improvements to the other process. Reordering rules that are learnt based on word alignments will potentially be better if the word alignment is improved. Word alignment tends to be better for similar languages, which can be simulated by applying reordering rules to the training data.

I used a reordering rule strategy similar to that in Elming (2008b), by using rule-induction learning on different levels of linguistic annotations, based on a dependency parse. I applied the rules to get a 1-best reordering of both the training data and the translation input. I performed experiments on English–German translation, comparing a baseline system without reordering to a system where reordering rules from 1 and 2 iterations were used. There were no improvements on the translation task over the baseline when using a decoder that allowed some reordering, and only minor improvements when using monotone decoding.

Rule-induction learning creates human-readable rules, and I could thus analyze the rules. The majority of the rules were linguistically motivated operations such as subject-verb inversion, and moving verbs to the end of sentences. The types of rules created in the two iterations were quite different. In iteration 1 there were many rules that moved verbs to the end of sentences, whereas there were mostly subject-verb inversion rules in iteration 2. Both these types of rules are useful for English–German reordering. This indicates that the iterative approach can aid in learning new types of rules, even though it had no overall effect on translation results. One reason for this could be that the rule-learner had a low recall.

4.3.5 Paper 5

Sara Stymne. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 28–34. Avignon, France

In this paper I investigated the use of clustered word classes for preordering, and compared it to several different other tagsets for English–German translation. I also used preordering based on clustered word classes for translation from Haitian Creole, a language for which no POS-taggers are available.

The work is based on a POS-based reordering algorithm by Niehues and Kolss (2009), which extracts both short-range and long-range rules based on alignments. The rules are expressed as POS-patterns, possibly with wild cards for covering long-distance reordering, with an associated order for each POS-pattern. For input to the decoder I used lattices that captured many possible reorderings. I used this algorithm both with a standard POS-tagset, and with several parsing-based tagsets, capturing dependencies and shallow syntax. The sizes of the tagsets varied between 20–523. I also investigated if automatically clustered word classes (Och, 1999) could be used instead of standard tags.

For English–German translation both the standard tagsets and the clustered tagsets gave improvements over a baseline without reorderings, both when the training data was reordered, and with the original order maintained. The systems with clustered word classes had slightly lower scores than some of the systems with standard tags. There were no consistent differences between the standard tagsets or between systems with a different number of word class clusters. For translation from Haitian Creole to English, the method with clustered word classes allowed us to use preordering, which is not possible with other annotations, since there is no available POS-tagger. Also for this language pair, which only has a moderate amount of reordering, there were consistent improvements over the baseline when word class reordering was used.

4.3.6 Paper 6

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and OOVs: Two problems for translation between German and English. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 183–188. Uppsala, Sweden

In this paper we investigated three strategies for translation between English and German: morphological processing, reordering for alignment, and OOV handling. The morphological processing included using a morphological sequence model and compound splitting, similarly to Paper 1. In this paper we also showed modest improvements using this strategy for translation from German to English. Reordering for alignment was applied using alignment-based reordering, as in Holmqvist et al. (2009). In this paper we also experimented with moving verbs to the end of sentences before alignment, with the hope that it would improve the alignment of verbs, which is problematic for translation between English and German. We found that on its own this alignment was not helpful compared to the baseline, but that it could be useful when combined with other alignments.

The main reason for including the paper in this thesis is for the discussion of OOV handling. Here we used two types of knowledge-light strategies, a preprocessing strategy for replacing OOVs with known equivalents, and a postprocessing step where number formatting was addressed. In the preprocessing step we performed four types of operations:

- Replace an OOV with a form with different casing (this could happen since we used truecasing in our system, not lowercasing).
- Stem the OOV and choose the most common known form for that word, if any.
- If the OOV was hyphenated, split on hyphens. If any of the parts were OOVs, repeat the above steps.
- Remove hyphens at the end of words (only for German, addresses coordinated compounds).

For English we found that many proper names were erroneously changed into other word types, and we thus excluded words starting with uppercase letters from the processing of English. These steps reduced the number of OOVs by 35.4% in German and 14.9% in English.

In the postprocessing step we changed the formatting of numbers to adhere to the formatting of the target language, when violated. In German a comma is used for decimals, and a period is used between thousands, whereas the usage in English is the opposite. We wrote simple regular expression-based rules for changing this.

Overall a relatively small number of words and sentences were affected by the two types of OOV processing, and the effect on MT metrics was minimal. The effect was, however, slightly positive in the majority of cases.

4.3.7 Paper 7

Sara Stymne. 2011c. Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 470–477. Edinburgh, Scotland

In this paper I reported results for translation of Haitian Creole SMS data to English. There were two conditions for the translation, the raw condition which used unmodified, raw data, and the clean condition, where development and test data had been cleaned by human annotators. The main contributions of the paper are the use of a cleaning model and a spell checking-based algorithm for OOV-replacement, which are summarized in Figure 4.2. It illustrates how PBSMT training were used both on word level for creating a cleaning model, and on character level for training weights for

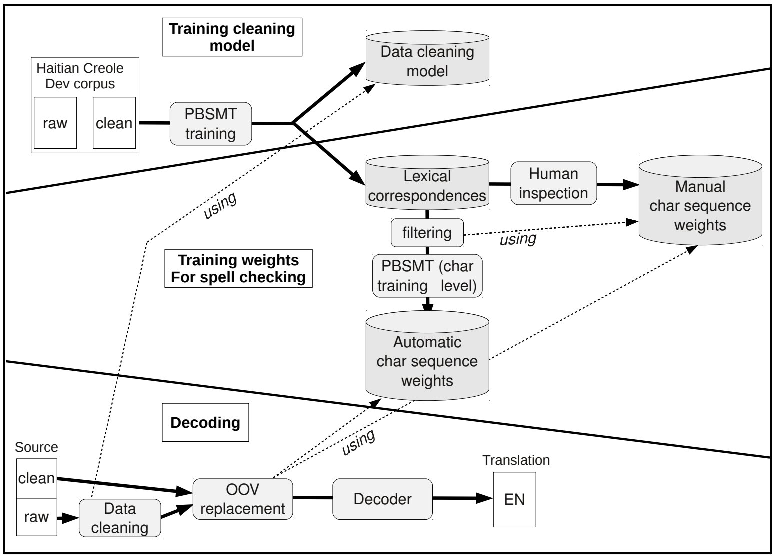

Figure 4.2: Overview of the training of the spell checking based OOV-replacement and cleaning model for translation of Haitian Creole in Paper 7.

OOV replacements, further described below. There is also some discussion on general corpus preparation, and of how to combine in- and out-of-domain data in the best way.

In the data cleaning step a small amount of manually cleaned data, 900 sentences from the development corpus, was used to train a PBSMT model that could translate from raw Haitian Creole to cleaner Haitian Creole. This model was applied to both the training data, and to the raw translation input. Despite the fact that it was trained on very little data, I found that it was useful for both the training data and translation input for the raw condition, but that it was not useful when the translation input had been cleaned.

For the OOV-replacement I applied a spell checking-based model to OOVs in the translation input, both performing a 1-best replacement and creating a lattice with the 3-best options. The spell checking algorithm allowed edit distance operations on character sequences, up to length 3, not only on single characters. Contrary to previous methods for spell checking-based replacement of OOVs, I allowed multiple edits in each string, and I used weighted edits, and also weights from a source-side language model. Allowing multiple multi-character edits was clearly useful since the corpus contained numerous spelling alternatives with several edits, such as *ayeropò/ayóport* (airport) and *pwogram/programme* (program).

Several weighting schemes were investigated. In the first scheme a subset of common substitutions was identified manually by inspecting a list of possible spelling alternatives that was created from raw and cleaned data. These common substitutions were given a uniform low cost. In the second scheme weights were trained automatically using character-based PBSMT trained on a list of spelling alternatives created using the first method. Both these two schemes were used to weigh individual edit operations, to arrive at a total score for each spelling alternative based on a modified Levenshtein distance. The word level scores were also used to rank candidates, and to weigh the lattice. For this purpose I also used a source side language model that evaluated alternatives in the source side context. On the translation task there was a large improvement when using OOV-lattices, regardless of which type of weights that was used. For 1-best substitutions the results were mostly close to the baseline.

5 Conclusions

In this thesis I have presented a number of text harmonization strategies, focusing on four areas: compounding, definiteness, reordering, and unknown words. For all these areas I have either developed new methods, or extended existing methods, mainly using pre- and postprocessing techniques. The term text harmonization has not been used much in this context before. I think it is a useful umbrella term for the kind of processing that aims at making two texts more similar in one or more ways, which has been shown to be effective for improving PBSMT, both in this thesis and in previous research.

Addressing translation into compounding languages, I developed several new methods. I showed how sequence models based on customized POS-tags can be used for improving the order of compound parts, and designed both new heuristics and learning-based methods for compound merging, that can produce novel compounds. Overall these methods led to relatively consistent small improvements over baseline systems without compound processing, and they also led to a higher number of compounds in the translations, which is desirable. Definiteness has not been addressed for SMT into the Scandinavian languages before. I designed a relatively simple preprocessing method that gave very good results for translation from English to Danish, and smaller but consistent improvements when ported to other language pairs. For reordering I extended previously proposed reordering algorithms in two ways, by iteration of alignment and reordering, and by the use of clustered word classes. While these methods did not lead to improvements over standard reordering systems, I showed that the iteration method led to the identification of new types of linguistically motivated rules, and the use of word classes allows preordering to be used for less-resourced languages like Haitian Creole. Finally I identified methods for replacing OOVs in the translation input. I applied light-weight techniques for English–German translation with modest improvements. I also used an elaborate spelling replacement method, with good results for Haitian Creole.

I used shallow linguistic knowledge, in the form of POS-tags in several ways. In preprocessing steps, POS-tags were used to guide compound splitting, definiteness processing, and reordering. In postprocessing it was used for compound merging. It was also used in factored decoding with sequence models on either customized POS-tags for improving compound coalescence, or morphologically enriched tags for improving agreement. The tags used for compound parts were customized tags, relevant to compound formation. I believe this type of technique could be used also for other phenomena, such as improving the order of split morphs, for morphological splitting. In Paper 6 I compared POS-tags to parser-based tagging schemes, without finding any particular advantage of parsing-based tags. This is not the first work where POS-tags have been used in some way to extend and improve PBSMT, it rather adds to the overwhelming evidence that shallow linguistics can indeed be useful for improving PBSMT.

The standard PBSMT methods are language independent, but better suited for languages with similar structure. The methods I used for compounding and definiteness are language dependent, but as shown in the thesis, they work for a number of language pairs. The only language specific information used in the compound processing algorithm is the inventory of allowed compound forms for each language and the adaption of POS-tags for each tagset. It was, however, important with careful modification of the strategy for definiteness processing when adapting it to new languages. In this work the English or Italian source side was transformed to resemble the Scandinavian NP structure. This can be contrasted to previous work on English–Arabic translation, where instead the Arabic side, which marks definiteness with a prefix, has been transformed to resemble English (e.g., Lee, 2004; Badr et al., 2008). A fuller investigation into these options should be useful. I also believe that the proposed methods could be extended to even more language pairs. The methods used for reordering and OOV-replacement are largely language independent, even though they are only evaluated on two language pairs. It seems that it is advantageous to use some prior knowledge of the language pairs treated, however. With the use of text harmonization strategies, it is possible to better exploit the strengths of PBSMT by training models on harmonized texts.

The spell checking techniques developed were only applied to a very noisy SMS corpus, for which they were useful. But there is reason to believe that it could also be used for other corpora, seeing that previous work have shown some improvements using simpler spell checking techniques on such corpora. The current weight training algorithm used some human knowledge for assigning the initial weights. I believe that it is possible to assign weights automatically by iteratively identifying spelling alternatives, and refine the weights based on those. Weights can most likely be initialized uniformly in such an approach. If such a strategy works, it would be completely unsupervised. An issue to be further addressed is how to weigh replacement candidates in the best way, and it would also be desirable to set thresholds in some more principled way. Other avenues for future work is to apply the spell-checking algorithm in other ways, for instance by adding entries to the phrase-table (Habash, 2008), or by using it to normalize spelling in the corpus (Hewavitharana et al., 2011). It could also be combined with other OOV-handling strategies, such as the light-weight strategies in Paper 6.

There are several issues in connection with PBSMT systems that are not discussed much in the thesis. One such issue is casing, especially the task of restoring the correct capitalization in postprocessing. It is an orthogonal problem to the issues discussed in the thesis, but for a full MT system it is still important to handle, and there has been some research in this area (e.g., Wang et al., 2006). In my research I suggest one approach for noisy corpora in Paper 7, by lowercasing the source, keeping the target true-cased, and filtering out sentences with much upper-casing from frequency calculations. More experimentation needs to be done to really show the usefulness of these choices, but preliminary unpublished experiments show that the results are better with the mixed casing strategy, than by using a lowercased corpus. A risk of using truecasing instead of lowercasing is that of raising the number of OOVs, which is addressed in Paper 6. Another issue for re-casing is for translation into German with systems with compound processing, in which novel nouns are formed, which need an initial upper-cased letter. These words might not be handled by methods that rely on corpus frequencies for recasing. In previous work I have suggested using POS-tags, which are available in the factored models, to make capitalization decisions (Stymne, 2009a).

In a recent paper, Clark et al. (2011) brought up the issue of optimizer instability, which though it was relatively well-known, has not been addressed much. The issue is that the standard algorithm for optimizing the weights of an SMT system, minimum error-rate training, tends to be unstable, and give varying results. To address this, Clark et al. (2011) recommend using several optimizer runs in combination with hypothesis testing using approximate randomization. I used approximate randomization for hypothesis testing in most of the papers in this thesis, but I did mostly not use several optimizer runs. What I did was to run a high number of experiments with slightly different settings for each type of approach. If all, or a large majority, of such experiments point in the same direction, it is very likely that the results are not due to variations in optimization. I thus think that the results in this thesis are valid despite the lack of repetition of identical experiments with multiple optimizer runs.

The main type of evaluation used in the thesis is automatic metrics, like Bleu (Papineni et al., 2002), which compares the translation output to one or a few human translations. Such metrics have limitations. In the case of only one reference translation, as used in this thesis, only one valid translation option is taken into account. Furthermore, metrics tend to give better results for languages with little morphology, such as English, for which the agreement with human evaluation tend to be better than for morphologically complex languages such as German (see e.g., Callison-Burch et al., 2011). Metrics like Bleu are also problematic for compounding languages, since a long compound corresponds to several words in a language like English. Even if it is correct, a compound word will only count as a unigram match in a compounding language, whereas it will count as a longer n-gram match in English, and thus count proportionally more in the final score. To alleviate these problems, two major strategies were used in the thesis. The first was to use several metrics, where the metrics at least to some extent measures different aspects of a translation. Many issues are common for all these metrics, however. For the work on compounding, I also used several additional analysis methods, such as counting compounds in the output, and judging their quality. These investigations showed that the compounding processing strategies proposed do lead to a higher number of compounds than the baseline system, with a good ordering of split compounds. For the other areas, the amount of additional analysis was limited. A thorough error analysis, to investigate particular strengths and weaknesses of the proposed systems would be valuable.

One issue that I saw for compound processing and definiteness processing is that the results varied somewhat for the two corpora used, Europarl and the automotive corpus. This could be due to the size of the corpora, Europarl is much larger, but it could also be due to the nature of Europarl, which is quite diverse, and which has different source languages for different parts, which means that the parallel texts are often translations from a third language. Much of the MT research on European languages is performed using the Europarl corpus, and it is thus an important issue that techniques that do not work on Europarl might work well on other corpora, and vice versa.

In this thesis, the focus has been solely on PBSMT. The techniques that were used here could be used in connection with hierarchical or syntactic SMT as well, however. There have been some studies indicating the usefulness of preprocessing also for those methods (e.g., Xu et al., 2009; Wang et al., 2010). Another direction for future work is to extend the current work to new language pairs and language phenomena. In this work English is mostly used as one of the languages; it would, for instance, also be interesting to see the effect of using compound processing on both sides when translating between two compounding languages.

In summary, I have shown several ways in which text harmonization techniques can be used to improve PBSMT. For the areas of compound processing for translation into compounding languages and definiteness for translation into Scandinavian languages I have developed several new methods that are superior to previous proposals. For reordering I have extended previously suggested preordering strategies in two ways, by iterating word alignment and reordering rule learning, and by using clustered word classes. Finally I have investigated knowledge-light methods for OOV-replacement, including a more elaborate spelling model than in previous research in SMT.

Bibliography

Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Columbus, Ohio, USA.

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 529–536. Sydney, Australia.

Enrique Alfonseca, Slaven Bilac, and Stefan Pharies. 2008a. Decompounding query keywords from compounding languages. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies, Short papers*, pages 253–256. Columbus, Ohio, USA.

Enrique Alfonseca, Slaven Bilac, and Stefan Pharies. 2008b. German decompounding in a difficult corpus. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 128–139. Haifa, Israel.

Jeffrey Allen. 1998. Lexical variation in Haitian Creole and orthographic issues for machine translation (MT) and optical character recognition (OCR) applications. In *The AMTA Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*. Langhorne, Pennsylvania, USA.

Hiyan Alshawi, Adam L. Buchsbaum, and Fei Xia. 1997. A comparison of head transducers and transfer for a limited domain translation application. In *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 360–365. Madrid, Spain.

Karunesh Arora, Michael Paul, and Eiichiro Sumita. 2008. Translation of unknown words in phrase-based statistical machine translation for languages of rich morphology. In *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU-2008)*, pages 70–75. Hanoi, Vietnam.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the*

46th Annual Meeting of the ACL: Human Language Technologies, pages 763–770. Columbus, Ohio, USA.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Poster session*, pages 33–40. Sydney, Australia.

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies, Short papers*, pages 153–156. Columbus, Ohio, USA.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*. Ann Arbor, Michigan, USA.

Laurie Bauer. 1983. *English Word Formation*. Cambridge, UK: Cambridge UP.

Laurie Bauer. 1998. When is a sequence of two nouns a compound in English? *English Language and Linguistics*, 2(1), pages 65–86.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, Andrew S. Kehler, and Robert L. Mercer. 1996. Language translation apparatus and method of using context-based translation models. United States patent, patent number 5510981, April.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the NAACL*, pages 412–419. Los Angeles, California, USA.

André Berton, Pablo Fetter, and Peter Regel-Brietzmann. 1996. Compound words in large-vocabulary German speech recognition systems. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 1165–1168. Philadelphia, Pennsylvania, USA.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16. Prague, Czech Republic.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754. Honolulu, Hawaii, USA.

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 129–135. Tokyo, Japan.

Ondřej Bojar. 2007. English-to-Czech factored machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239. Prague, Czech Republic.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 286–293. Hong Kong.

Benny Brodda. 1979. Något om de svenska ordens fonotax och morfotax: Iakttagelse med utgångspunkt från experiment med automatisk morfologisk analys. In *PILUS nr 38*. Inst. för lingvistik, Stockholms universitet, Sweden.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), pages 263–311.

Ralf D. Brown. 2002. Corpus-driven splitting of compound words. In *Proceedings of the 9th International Conference of Theoretical and Methodological Issues in Machine Translation*, pages 12–21. Keihanna, Japan.

Lars Bungum and Stephan Oepen. 2009. Automatic translation of Norwegian noun compounds. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 136–143. Barcelona, Spain.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Prague, Czech Republic.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Columbus, Ohio, USA.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28. Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Edinburgh, Scotland.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. In *Proceedings of the 2006 Human Language Technology Conference of the NAACL*, pages 17–24. New York City, New York, USA.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of the 11th Conference of the EACL*, pages 249–256. Trento, Italy.

Johan Carlberger, Rickard Domeij, Viggo Kann, and Ola Knutsson. 2005. The development and performance of a grammar checker for Swedish: A language engineering perspective. In Ola Knutsson. 2005. *Developing and Evaluating Language Tools for Writers and Learners of Swedish.* Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm, Sweden.

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 29, pages 815–832.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183. Uppsala, Sweden.

Xavier Carreras and Michael Collins. 2009. Non-projective parsing for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 200–209. Singapore.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4), pages 359–393.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270. Ann Arbor, Michigan, USA.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition*, 10, pages 157–174.

Kenneth W. Church and William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing*, 1, pages 93–103.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 176–181. Portland, Oregon, USA.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540. Ann Arbor, Michigan, USA.

Bernard Comrie. 1981. *Language Universals and Linguistic Typology*. Oxford: Blackwell.

Marta R. Costa-jussà and José A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 70–76. Sydney, Australia.

Marta R. Costa-jussà and José A. R. Fonollosa. 2007. Analysis of statistical and morphological classes to generate weigthed reordering hypotheses on a statistical machine translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 171–176. Prague, Czech Republic.

Josep M. Crego and Nizar Habash. 2008. Using shallow syntax information to improve word alignment and reordering for SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61. Columbus, Ohio.

Josep M. Crego and José B. Mariño. 2006. Integration of POStag-based source reordering into SMT decoding by an extended search graph. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 29–36. Cambridge, Massachusetts, USA.

Josep M. Crego and José B. Mariño. 2007. Syntax-enhanced n-gram-based SMT. In *Proceedings of MT Summit XI*, pages 111–118. Copenhagen, Denmark.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140. Trento, Italy.

Östen Dahl. 2003. Definite articles in Scandinavian: Competing grammaticalization processes in standard and non-standard varieties. In Bernd Kortmann, editor, *Dialect Grammar from a Cross-Linguistic Perspective*, pages 147–180. Berlin: Mouton de Gruyter.

Fred Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3), pages 659–664.

Arthur E. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1), pages 1–38.

Steve DeNeefe, Ulf Hermjakob, and Kevin Knight. 2008. Overcoming vocabulary sparsity in MT using lattices. In *Proceedings of the 8th Conference* *of the Association for Machine Translation in the Americas*, pages 89–96. Waikiki, Hawaii, USA.

Etienne Denoual. 2007. Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of MT Summit XI*, pages 135–141. Copenhagen, Denmark.

Mona Diab, Mahmoud Ghoneim, and Nizar Habash. 2007. Arabic diacritization in the context of statistical machine translation. In *Proceedings of MT Summit XI*, pages 143–149. Copenhagen, Denmark.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Proceedings of the Second International Conference on Human Language Technology*, pages 228–231. San Diego, California, USA.

Stephen Doherty and Sharon O'Brien. 2009. Can MT output be evaluated through eye tracking? In *Proceedings of MT Summit XII*, pages 214–221. Ottawa, Ontario, Canada.

Wolfgang U. Dressler. 2006. Compound types. In Gary Libben and Gonia Jarema, editors, *Representation and Processing of Compound Words*, pages 23–44. Oxford, UK: Oxford UP.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194. Columbus, Ohio, USA.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the NAACL*, pages 406–414. Boulder, Colorado, USA.

Chris Dyer. 2010. *A Formal Model of Ambiguity and its Applications in Machine Translation*. Ph.D. thesis, University of Maryland, USA.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies*, pages 1012–1020. Columbus, Ohio, USA.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2008. Communicating unknown words in machine translation. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Anas El Isbihani, Shahram Khadivi, Oliver Bender, and Hermann Ney. 2006. Morpho-syntactic Arabic preprocessing for Arabic to English statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 15–22. New York City, New York, USA.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in English to Turkish statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 7–14. New York City, New York, USA.

İlknur Durgar El-Kahlout and François Yvon. 2010. The pay-offs of preprocessing for German-English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 251–258. Paris, France.

Ahmed El Kholy and Nizar Habash. 2010. Techniques for Arabic morphological detokenization and orthographic denormalization. In *LREC 2010 Workshop on Language Resources and Human Language Technology for Semitic Languages*, pages 45–51. Valletta, Malta.

Jakob Elming. 2006. Transformation-based correction of rule-based MT. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, pages 219–226. Oslo, Norway.

Jakob Elming. 2008a. *Syntactic Reordering in Statistical Machine Translation*. Ph.D. thesis, Copenhagen Business School, Denmark.

Jakob Elming. 2008b. Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 46–54. Columbus, Ohio, USA.

Alexander Fraser. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 115–119. Athens, Greece.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of the 13th Conference of the EACL*, pages 664–674. Avignon, France.

Karin Friberg. 2007. Decomposing Swedish compounds using memory-based learning. In *Proceedings of the 16th Nordic Conference on Computational Linguistics (NODALIDA'07)*, pages 224–230. Tartu, Estonia.

Fabienne Fritzinger and Alexander Fraser. 2010. How to avoid burning ducks: Combining linguistic analysis and corpus statistics for german compound processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 224–234. Uppsala, Sweden.

Masaru Fuji. 1999. Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of MT Summit VII*, pages 285–289. Singapore.

William A. Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), pages 75–102.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntatic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 961–968. Sydney, Australia.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Honolulu, Hawaii, USA.

Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the NAACL*, pages 966–974. Los Angeles, California, USA.

Nikesh Garera and David Yarowsky. 2008. Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 403–410. Hyderabad, India.

Barbara Gawrońska, Anders Nordner, Christer Johansson, and Caroline Willners. 1994. Interpreting compounds for machine translation. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 45–50. Kyoto, Japan.

Barbara Gawrońska-Werngren. 1990. ”Translation great problem” – on the problem of inserting articles when translating from Russian into Swedish. In *Coling-90: Papers presented to the 13th International Conference on Computational Linguistics, vol. 2*, pages 133–138. Helsinki, Finland.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 376–384. Beijing, China.

Jesús Giménez and Lluìs Márquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198. Columbus, Ohio, USA.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint*

Conference on Natural Language Processing of the AFNLP, Short papers, pages 73–76. Boulder, Colorado, USA.

John Goldsmith and Tom Reutter. 1998. Automatic collection and analysis of German compounds. In *Proceedings of the Coling-ACL Workshop on the Computational Treatment of Nominals Workshop*, pages 61–69. Montreal, Quebec, Canada.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Human Language Technology Conference and the conference on Empirical Methods in Natural Language Processing*, pages 676–683. Vancouver, British Columbia, Canada.

Cyril Goutte, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 502–509. Barcelona, Spain.

Jonathan Graehl and Kevin Knight. 2004. Training tree transducers. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL*, pages 105–112. Boston, Massachusetts, USA.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of MT Summit XI*, pages 215–222. Copenhagen, Denmark.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies, Short papers*, pages 57–60. Columbus, Ohio, USA.

Nizar Habash and Hayden Metsky. 2008. Automatic learning of morphological variations for handling out-of-vocabulary terms in Urdu-English machine translation. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pages 107–116. Waikiki, Hawaii, USA.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 2006 Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52. New York City, New York, USA.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378. Portland, Oregon, USA.

Jorge Hankamer and Line Mikkelsen. 2002. A morphological analysis of definite nouns in Danish. *Journal of Germanic Linguistics*, 14(2), pages 137–175.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Edinburgh, UK.

Turid Hedlund. 2002. Compounds in dictionary-based cross-language information retrieval. *Information Research*, 7(2). Available at `http://InformationR.net/ir/7-2/paper128.html` (visited March 29, 2012).

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies*, pages 389–397. Columbus, Ohio.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. CMU Haitian Creole-English translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 386–392. Edinburgh, Scotland.

Maria Holmqvist, Sara Stymne, and Lars Ahrenberg. 2007. Getting to know Moses: Initial experiments on German-English factored translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 181–184. Prague, Czech Republic.

Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 120–124. Athens, Greece.

Florian Holz and Chris Biemann. 2008. Unsupervised and knowledge-free learning of compound splits and periphrases. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 117–127. Haifa, Israel.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 144–151. Prague, Czech Republic.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. A syntax-directed translator with extended domain of locality. In *Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, pages 1–8. New York City, New York, USA.

Institut für Deutsche Sprache. 1998. Rechtschreibreform (Aktualisierte Ausgabe). IDS Sprachreport, Extra-Ausgabe Dezember 1998. Mannheim, Germany.

Harri Jäppinen and Leo Kulikov. 1991. Evaluation of machine translation systems: A system developer's viewpoint. In *Proceedings of the Evaluators' Forum*, pages 143–156. Les Rasses, Switzerland.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1009–1012. Philadelphia, Pennsylvania, USA.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the 12th National conference of the American Association for Artificial Intelligence*. Seattle, Washington, USA.

Catherine Kobus, Franois Yvon, and Graldine Damnati. 2008. Transcrire les SMS comme on reconnaît la parole. In *Actes de la Conférence sur le Traitement Automatique des Langues (TALN'08)*, pages 128–138. Avignon, France.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395. Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86. Phuket, Thailand.

Philipp Koehn. 2009. *Moses, a Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models, User Manual and Code Guide*. University of Edinburgh. Software Manual.

Philipp Koehn, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142. Columbus, Ohio, USA.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*. Pittsburgh, Pennsylvania, USA.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876. Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, pages 177–180. Prague, Czech Republic.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the EACL*, pages 187–193. Budapest, Hungary.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the NAACL*, pages 48–54. Edmonton, Alberta, Canada.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227. Prague, Czech Republic.

Dimitros Kokkinakis. 2001. *A Framework for the Acquisition of Lexical Knowledge: Description and Applications*. Ph.D. thesis, Göteborg University, Sweden.

Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2006. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 77–82. Kyoto, Japan.

Roland Kuhn, Denis Yuen, Michel Simard, Patrick Paul, George Foster, Eric Joanis, and Howard Johnson. 2006. Segment choice models: Feature-rich models for global distortion in statistical machine translation. In *Proceedings of the 2006 Human Language Technology Conference of the NAACL*, pages 25–32. New York City, New York, USA.

Sebastian Kürschner. 2003. *Von Volk-s-musik und Sport-ø-geist im Lemming-ø-land – Af folk-e-musik og sport-s-ånd i lemming-e-landet: Fugenelemente im Deutschen und Dänischen – eine kontrastive Studie zu einem Grenzfall der Morphologie*. Master's thesis, Albert-Ludwigs-Universität, Freiburg, Germany.

Stefan Langer. 1998. Zur Morphologie und Semantik von Nominalkomposita. In *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 83–97. Bonn, Germany.

Philippe Langlais and Alexandre Patry. 2007. Translating unknown words using analogical learning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 877–886. Prague, Czech Republic.

Martha Larson, Daniel Willett, Joachim Köhler, and Gerhard Rigoll. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, volume 3, pages 945–948. Beijing, China.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Prague, Czech Republic.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL: Short Papers*, pages 57–60. Boston, Massachusetts, USA.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8), pages 707–710.

William D. Lewis. 2010. Haitian Creole: how to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation.* Saint-Raphaël, France.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 720–727. Prague, Czech Republic.

Jin-Ji Li, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 190–196. Athens, Greece.

Gary Libben, Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84, pages 50–64.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 609–616. Sydney, Australia.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 976–985. Prague, Czech Republic.

Christopher Lyons. 1999. *Definiteness.* Cambridge: Cambridge UP.

Klaus Macherey, Andrew Dai, David Talbot, Ashok Popat, and Franz Och. 2011. Language-independent compound splitting with morphological operations. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 1395–1404. Portland, Oregon, USA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Pennsylvania, Pennsylvania, USA.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4), pages 527–549.

Montserrat Meya. 1990. Tenets for an interlingual representation of definite NPs. In *Coling-90: Papers presented to the 13th International Conference on Computational Linguistics, vol. 2*, pages 263–269. Helsinki, Finland.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 128–135. Prague, Czech Republic.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 791–799. Suntec, Singapore.

Hanne Moa. 2005. Compounds and other oddities in machine translation. In *Proceedings of the 15th Nordic Conference on Computational Linguistics (NODALIDA'05)*, pages 124–132. Joensuu, Finland.

Robert Munro. 2011. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 68–77. Portland, Oregon, USA.

Hwidong Na, Jin-Ji Li, Jungi Kim, and Jong-Hyeok Lee. 2009. Improving fluency by reordering target constituents using MST parser in English-to-Japanese phrase-based SMT. In *Proceedings of MT Summit XII*, pages 276–283. Ottawa, Ontario, Canada.

Jan Niehues, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe translation system for the EACL-WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 80–84. Athens, Greece.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 206–214. Athens, Greece.

Sonja Nießen. 2002. *Improving statistical machine translation using morpho-syntactic information*. Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule, Aachen, Germany.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 1081–1085. Saarbrücken, Germany.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2), pages 181–204.

Eric W. Noreen. 1998. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. New York: Wiley.

Diarmuid Ó Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*. Birmingham, UK.

Diarmuid Ó Séaghdha and Ann Copestake. 2009. Using lexical and relational similarity to classify semantic relations. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 621–629. Athens, Greece.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the EACL*, pages 71–76. Bergen, Norway.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 160–167. Sapporo, Japan.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL*, pages 161–168. Boston, Massachusetts, USA.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447. Hong Kong.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302. Philadelphia, Pennsylvania, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pages 19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28. College Park, Maryland, USA.

Lene Offersgaard, Claus Povlsen, Lisbeth Almsten, and Bente Maegaard. 2008. Domain specific MT in use. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 150–159. Hamburg, Germany.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, pages 80–87. Rochester, New York, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318. Philadelphia, Pennsylvania, USA.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2009. NICT@WMT09: Model adaptation and transliteration for Spanish-English SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 105–109. Athens, Greece.

Deana Pennell and Yang Liu. 2011. A character-level machine translation approach for normalization of SMS abbreviations. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 974–982. Chiang Mai, Thailand.

Maja Popović and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1585–1588. Lisbon, Portugal.

Maja Popović and Hermann Ney. 2006. POS-based reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1278–1283. Genoa, Italy.

Maja Popović and Hermann Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32. Athens, Greece.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of German compound words. In *Proceedings of FinTAL – 5th International Conference on Natural Language Processing*, pages 616–624. Turku, Finland: Springer Verlag, LNCS.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 271–279. Ann Arbor, Michigan, USA.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language.* London, England: Longman.

Ulrike Rackow, Ido Dagan, and Ulrike Schwall. 1992. Automatic translation of noun compounds. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 1249–1253. Nantes, France.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 800–808. Suntec, Singapore.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*, pages 57–64. Ann Arbor, Michigan, USA.

Kay Rottmann and Stephan Vogel. 2007. Word reordering in statistical machine translation with a POS-based distortion model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 171–180. Skövde, Sweden.

Markus Saers. 2011. *Translation as Linear Transduction: Models and Algorithms for Efficient Learning in Statistical Machine Translation.* Ph.D. thesis, Uppsala University, Uppsala, Sweden.

Yvonne Samuelsson. 2006. Nouns in statistical machine translation. Unpublished manuscript: Term paper, Statistical Machine Translation. Course given at Copenhagen Business School, Denmark.

Bambi Schiefflin and Rachelle Charlier Doucet. 1992. The 'real' Haitian Creole: Metalinguistics and orthographic choice. *Pragmatics*, 2(3), pages 427–443.

Anne Schiller. 2005. German compound analysis with wfsc. In *Proceedings of the Finite State Methods and Natural Language Processing*, pages 239–246. Helsinki, Finland: Springer Verlag, LNCS.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. Manchester, UK.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 777–784. Manchester, UK.

Ylva Hård af Segerstad. 2002. *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication.* Ph.D. thesis, Göteborg University, Sweden.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL*, pages 177–184. Boston, Massachusetts, USA.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4), pages 649–671.

Wade Shen, Richard Zens, Nicola Bertoldi, and Marcello Federico. 2006. The JHU workshop 2006 IWSLT system. In *Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation*, pages 59–63. Kyoto, Japan.

Stuart M. Shieber. 2007. Probabilistic synchronous tree-adjoining grammars for machine translation: The argument from bilingual dictionaries. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, pages 88–95. Rochester, New York, USA.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of the Human Language Technology Conference and the conference on Empirical Methods in Natural Language Processing*, pages 755–762. Vancouver, British Columbia, Canada.

Jonas Sjöbergh and Viggo Kann. 2004. Finding the correct interpretation of Swedish compounds, a statistical approach. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 899–902. Lisbon, Portugal.

David Smith and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 23–30. New York City, New York, USA.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of nonstandard words. *Computer Speech and Language*, 15(3), pages 287–333.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904. Denver, Colorado, USA.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In *Proceedings of GoTAL – 6th International Conference on Natural Language Processing*, pages 464–475. Gothenburg, Sweden: Springer Verlag, LNCS/LNAI.

Sara Stymne. 2009a. A comparison of merging strategies for translation of German compounds. In *Proceedings of the EACL 2009 Student Research Workshop*, pages 61–69. Athens, Greece.

Sara Stymne. 2009b. *Compound processing for phrase-based statistical machine translation.* Licentiate thesis, Linköping University, Sweden.

Sara Stymne. 2009c. Definite noun phrases in statistical machine translation into Danish. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 4–9. Odense, Denmark.

Sara Stymne. 2011a. Definite noun phrases in statistical machine translation into Scandinavian languages. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 289–296. Leuven, Belgium.

Sara Stymne. 2011b. Iterative reordering and word alignment for statistical MT. In *Proceedings of the 18th Nordic Conference on Computational Linguistics (NODALIDA'11)*, pages 315–318. Riga, Latvia.

Sara Stymne. 2011c. Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 470–477. Edinburgh, Scotland.

Sara Stymne. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 28–34. Avignon, France.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 2175–2181. Valetta, Malta.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

Sara Stymne and Nicola Cancedda. 2011. Productive generation of compound words in statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 250–260. Edinburgh, Scotland.

Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2012a. Generation of compound words for statistical machine translation into compounding languages. Submitted manuscript.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012b. Eye tracking as a tool for machine translation error analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138. Columbus, Ohio, USA.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and OOVs: Two problems for translation between German and English. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 183–188. Uppsala, Sweden.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 969–976. Sydney, Australia.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24. Sapporo, Japan.

Paso Tapanainen and Timo Järvinen. 1997. A nonprojective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71. Washington, DC, USA.

Olof Thorell. 1981. *Svensk ordbildningslära*. Stockholm, Sweden: Esselte Studium.

Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference* *of the European Chapter of the Association for Computational Linguistics*, pages 141–151. Avignon, France.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the 2004 Human Language Technology Conference of the NAACL: Short Papers*, pages 101–104. Boston, Massachusetts, USA.

Christoph Tillmann. 2008. A rule-driven dynamic programming decoder for statistical MT. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, pages 37–45. Columbus, Ohio.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2667–2670. Rhodes, Greece.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technologies*, pages 514–522. Columbus, Ohio.

Albert Valdmann. 1970. *Basic Course in Haitian Creole*, volume 5 of *Language Science Monographs*. Bloomington: Indiana University.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39. Prague, Czech Republic.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702. Genoa, Italy.

Sami Virpioja, Jaako J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of MT Summit XI*, pages 491–498. Copenhagen, Denmark.

Stephan Vogel, Hermann Ney, and Christoph Tillman. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841. Copenhagen, Denmark.

Stephan Vogel, Sonja Nießen, and Hermann Ney. 2000. Automatic extrapolation of human assessment of translation quality. In *Proceedings of the Workshop on the Evaluation of Machine Translation at LREC'2000*, pages 35–39. Athens, Greece.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745. Prague, Czech Republic.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 1–8. New York City, New York, USA.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Restructuring, re-labeling, and re-aligning for syntax-based machine translation. *Computational Linguistics*, 36(2), pages 247–277.

Ian H. Witten and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4), pages 1085–1094.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3), pages 377–404.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514. Geneva, Switzerland.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253. Boulder, Colorado.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 303–310. Philadelphia, Pennsylvania, USA.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the 11th Conference of the EACL*, pages 41–48. Trento, Italy.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 205–211. Geneva, Switzerland.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *Proceedings of the German Conference on Artificial Intelligence (KI 2002)*, pages 18–32. Aachen, Germany.

Min Zhang, Hongfei Jiang, Ai Ti Aw, Jun Sun, Sheng Li, and Chew Lim Tan. 2007a. A tree-to-tree alignment-based model for statistical machine translation. In *Proceedings of MT Summit XI*, pages 535–542. Copenhagen, Denmark.

Ruiqiang Zhang and Eiichiro Sumita. 2008. Chinese unknown word translation by subword re-segmentation. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 225–232. Hyderabad, India.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: how much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 2051–2054. Lisbon, Portugal.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007b. Improved chunk-level reordering for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 21–28. Trento, Italy.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141. New York City; New York, USA.

Simon Zwarts and Mark Dras. 2008. Morphosyntactic target language matching in statistical machine translation. In *Proceedings of the Australasian Language Technology Workshop 2008*, pages 169–177. Hobart, Australia.

Department of Computer and Information Science
Linköpings universitet

Dissertations

Linköping Studies in Science and Technology
Linköping Studies in Arts and Science
Linköping Studies in Statistics
Linköpings Studies in Informatics

Linköping Studies in Science and Technology

No 14 **Anders Haraldsson:** A Program Manipulation System Based on Partial Evaluation, 1977, ISBN 91-7372-144-1.

No 17 **Bengt Magnhagen:** Probability Based Verification of Time Margins in Digital Designs, 1977, ISBN 91-7372-157-3.

No 18 **Mats Cedwall**: Semantisk analys av process-beskrivningar i naturligt språk, 1977, ISBN 91- 7372-168-9.

No 22 **Jaak Urmi:** A Machine Independent LISP Compiler and its Implications for Ideal Hardware, 1978, ISBN 91-7372-188-3.

No 33 **Tore Risch:** Compilation of Multiple File Queries in a Meta-Database System 1978, ISBN 91- 7372-232-4.

No 51 **Erland Jungert:** Synthesizing Database Structures from a User Oriented Data Model, 1980, ISBN 91-7372-387-8.

No 54 **Sture Hägglund:** Contributions to the Development of Methods and Tools for Interactive Design of Applications Software, 1980, ISBN 91-7372-404-1.

No 55 **Pär Emanuelson:** Performance Enhancement in a Well-Structured Pattern Matcher through Partial Evaluation, 1980, ISBN 91-7372-403-3.

No 58 **Bengt Johnsson, Bertil Andersson:** The Human-Computer Interface in Commercial Systems, 1981, ISBN 91-7372-414-9.

No 69 **H. Jan Komorowski:** A Specification of an Abstract Prolog Machine and its Application to Partial Evaluation, 1981, ISBN 91-7372-479-3.

No 71 **René Reboh:** Knowledge Engineering Techniques and Tools for Expert Systems, 1981, ISBN 91-7372-489-0.

No 77 **Östen Oskarsson:** Mechanisms of Modifiability in large Software Systems, 1982, ISBN 91- 7372-527-7.

No 94 **Hans Lunell:** Code Generator Writing Systems, 1983, ISBN 91-7372-652-4.

No 97 **Andrzej Lingas:** Advances in Minimum Weight Triangulation, 1983, ISBN 91-7372-660-5.

No 109 **Peter Fritzson:** Towards a Distributed Programming Environment based on Incremental Compilation, 1984, ISBN 91-7372-801-2.

No 111 **Erik Tengvald:** The Design of Expert Planning Systems. An Experimental Operations Planning System for Turning, 1984, ISBN 91-7372- 805-5.

No 155 **Christos Levcopoulos:** Heuristics for Minimum Decompositions of Polygons, 1987, ISBN 91-7870-133-3.

No 165 **James W. Goodwin:** A Theory and System for Non-Monotonic Reasoning, 1987, ISBN 91-7870-183-X.

No 170 **Zebo Peng:** A Formal Methodology for Automated Synthesis of VLSI Systems, 1987, ISBN 91-7870-225-9.

No 174 **Johan Fagerström:** A Paradigm and System for Design of Distributed Systems, 1988, ISBN 91-7870-301-8.

No 192 **Dimiter Driankov:** Towards a Many Valued Logic of Quantified Belief, 1988, ISBN 91-7870-374-3.

No 213 **Lin Padgham:** Non-Monotonic Inheritance for an Object Oriented Knowledge Base, 1989, ISBN 91-7870-485-5.

No 214 **Tony Larsson:** A Formal Hardware Description and Verification Method, 1989, ISBN 91-7870-517-7.

No 221 **Michael Reinfrank:** Fundamentals and Logical Foundations of Truth Maintenance, 1989, ISBN 91-7870-546-0.

No 239 **Jonas Löwgren:** Knowledge-Based Design Support and Discourse Management in User Interface Management Systems, 1991, ISBN 91-7870-720-X.

No 244 **Henrik Eriksson:** Meta-Tool Support for Knowledge Acquisition, 1991, ISBN 91-7870-746-3.

No 252 **Peter Eklund:** An Epistemic Approach to Interactive Design in Multiple Inheritance Hierarchies, 1991, ISBN 91-7870-784-6.

No 258 **Patrick Doherty:** NML3 - A Non-Monotonic Formalism with Explicit Defaults, 1991, ISBN 91-7870-816-8.

No 260 **Nahid Shahmehri:** Generalized Algorithmic Debugging, 1991, ISBN 91-7870-828-1.

No 264 **Nils Dahlbäck:** Representation of Discourse-Cognitive and Computational Aspects, 1992, ISBN 91-7870-850-8.

No 265 **Ulf Nilsson:** Abstract Interpretations and Abstract Machines: Contributions to a Methodology for the Implementation of Logic Programs, 1992, ISBN 91-7870-858-3.

No 270 **Ralph Rönnquist:** Theory and Practice of Tense-bound Object References, 1992, ISBN 91-7870-873-7.

No 273 **Björn Fjellborg:** Pipeline Extraction for VLSI Data Path Synthesis, 1992, ISBN 91-7870-880-X.

No 276 **Staffan Bonnier:** A Formal Basis for Horn Clause Logic with External Polymorphic Functions, 1992, ISBN 91-7870-896-6.

No 277 **Kristian Sandahl:** Developing Knowledge Management Systems with an Active Expert Methodology, 1992, ISBN 91-7870-897-4.

No 281 **Christer Bäckström:** Computational Complexity of Reasoning about Plans, 1992, ISBN 91-7870-979-2.

No 292 **Mats Wirén:** Studies in Incremental Natural Language Analysis, 1992, ISBN 91-7871-027-8.

No 297 **Mariam Kamkar:** Interprocedural Dynamic Slicing with Applications to Debugging and Testing, 1993, ISBN 91-7871-065-0.

No 302 **Tingting Zhang:** A Study in Diagnosis Using Classification and Defaults, 1993, ISBN 91-7871-078-2

No 312 **Arne Jönsson:** Dialogue Management for Natural Language Interfaces - An Empirical Approach, 1993, ISBN 91-7871-110-X.

No 338 **Simin Nadjm-Tehrani:** Reactive Systems in Physical Environments: Compositional Modelling and Framework for Verification, 1994, ISBN 91-7871-237-8.

No 371 **Bengt Savén:** Business Models for Decision Support and Learning. A Study of Discrete-Event Manufacturing Simulation at Asea/ ABB 1968-1993, 1995, ISBN 91-7871-494-X.

No 375 **Ulf Söderman:** Conceptual Modelling of Mode Switching Physical Systems, 1995, ISBN 91-7871-516-4.

No 383 **Andreas Kågedal:** Exploiting Groundness in Logic Programs, 1995, ISBN 91-7871-538-5.

No 396 **George Fodor:** Ontological Control, Description, Identification and Recovery from Problematic Control Situations, 1995, ISBN 91-7871-603-9.

No 413 **Mikael Pettersson:** Compiling Natural Semantics, 1995, ISBN 91-7871-641-1.

No 414 **Xinli Gu:** RT Level Testability Improvement by Testability Analysis and Transformations, 1996, ISBN 91-7871-654-3.

No 416 **Hua Shu:** Distributed Default Reasoning, 1996, ISBN 91-7871-665-9.

No 429 **Jaime Villegas:** Simulation Supported Industrial Training from an Organisational Learning Perspective - Development and Evaluation of the SSIT Method, 1996, ISBN 91-7871-700-0.

No 431 **Peter Jonsson:** Studies in Action Planning: Algorithms and Complexity, 1996, ISBN 91-7871-704-3.

No 437 **Johan Boye:** Directional Types in Logic Programming, 1996, ISBN 91-7871-725-6.

No 439 **Cecilia Sjöberg:** Activities, Voices and Arenas: Participatory Design in Practice, 1996, ISBN 91-7871-728-0.

No 448 **Patrick Lambrix:** Part-Whole Reasoning in Description Logics, 1996, ISBN 91-7871-820-1.

No 452 **Kjell Orsborn:** On Extensible and Object-Relational Database Technology for Finite Element Analysis Applications, 1996, ISBN 91-7871-827-9.

No 459 **Olof Johansson:** Development Environments for Complex Product Models, 1996, ISBN 91-7871-855-4.

No 461 **Lena Strömbäck:** User-Defined Constructions in Unification-Based Formalisms, 1997, ISBN 91-7871-857-0.

No 462 **Lars Degerstedt:** Tabulation-based Logic Programming: A Multi-Level View of Query Answering, 1996, ISBN 91-7871-858-9.

No 475 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av hur ekonomiska styrsystem utformas och används efter företagsförvärv, 1997, ISBN 91-7871-914-3.

No 480 **Mikael Lindvall:** An Empirical Study of Requirements-Driven Impact Analysis in Object-Oriented Software Evolution, 1997, ISBN 91-7871-927-5.

No 485 **Göran Forslund:** Opinion-Based Systems: The Cooperative Perspective on Knowledge-Based Decision Support, 1997, ISBN 91-7871-938-0.

No 494 **Martin Sköld:** Active Database Management Systems for Monitoring and Control, 1997, ISBN 91-7219-002-7.

No 495 **Hans Olsén:** Automatic Verification of Petri Nets in a CLP framework, 1997, ISBN 91-7219-011-6.

No 498 **Thomas Drakengren:** Algorithms and Complexity for Temporal and Spatial Formalisms, 1997, ISBN 91-7219-019-1.

No 502 **Jakob Axelsson:** Analysis and Synthesis of Heterogeneous Real-Time Systems, 1997, ISBN 91-7219-035-3.

No 503 **Johan Ringström:** Compiler Generation for Data-Parallel Programming Languages from Two-Level Semantics Specifications, 1997, ISBN 91-7219-045-0.

No 512 **Anna Moberg:** Närhet och distans - Studier av kommunikationsmönster i satellitkontor och flexibla kontor, 1997, ISBN 91-7219-119-8.

No 520 **Mikael Ronström:** Design and Modelling of a Parallel Data Server for Telecom Applications, 1998, ISBN 91-7219-169-4.

No 522 **Niclas Ohlsson:** Towards Effective Fault Prevention - An Empirical Study in Software Engineering, 1998, ISBN 91-7219-176-7.

No 526 **Joachim Karlsson:** A Systematic Approach for Prioritizing Software Requirements, 1998, ISBN 91-7219-184-8.

No 530 **Henrik Nilsson:** Declarative Debugging for Lazy Functional Languages, 1998, ISBN 91-7219-197-x.

No 555 **Jonas Hallberg:** Timing Issues in High-Level Synthesis, 1998, ISBN 91-7219-369-7.

No 561 **Ling Lin:** Management of 1-D Sequence Data - From Discrete to Continuous, 1999, ISBN 91-7219-402-2.

No 563 **Eva L Ragnemalm:** Student Modelling based on Collaborative Dialogue with a Learning Companion, 1999, ISBN 91-7219-412-X.

No 567 **Jörgen Lindström:** Does Distance matter? On geographical dispersion in organisations, 1999, ISBN 91-7219-439-1.

No 582 **Vanja Josifovski:** Design, Implementation and Evaluation of a Distributed Mediator System for Data Integration, 1999, ISBN 91-7219-482-0.

No 589 **Rita Kovordányi:** Modeling and Simulating Inhibitory Mechanisms in Mental Image Reinterpretation - Towards Cooperative Human-Computer Creativity, 1999, ISBN 91-7219-506-1.

No 592 **Mikael Ericsson:** Supporting the Use of Design Knowledge - An Assessment of Commenting Agents, 1999, ISBN 91-7219-532-0.

No 593 **Lars Karlsson:** Actions, Interactions and Narratives, 1999, ISBN 91-7219-534-7.

No 594 **C. G. Mikael Johansson:** Social and Organizational Aspects of Requirements Engineering Methods - A practice-oriented approach, 1999, ISBN 91-7219-541-X.

No 595 **Jörgen Hansson:** Value-Driven Multi-Class Overload Management in Real-Time Database Systems, 1999, ISBN 91-7219-542-8.

No 596 **Niklas Hallberg:** Incorporating User Values in the Design of Information Systems and Services in the Public Sector: A Methods Approach, 1999, ISBN 91-7219-543-6.

No 597 **Vivian Vimarlund:** An Economic Perspective on the Analysis of Impacts of Information Technology: From Case Studies in Health-Care towards General Models and Theories, 1999, ISBN 91-7219-544-4.

No 598 **Johan Jenvald:** Methods and Tools in Computer-Supported Taskforce Training, 1999, ISBN 91-7219-547-9.

No 607 **Magnus Merkel:** Understanding and enhancing translation by parallel text processing, 1999, ISBN 91-7219-614-9.

No 611 **Silvia Coradeschi:** Anchoring symbols to sensory data, 1999, ISBN 91-7219-623-8.

No 613 **Man Lin:** Analysis and Synthesis of Reactive Systems: A Generic Layered Architecture Perspective, 1999, ISBN 91-7219-630-0.

No 618 **Jimmy Tjäder:** Systemimplementering i praktiken - En studie av logiker i fyra projekt, 1999, ISBN 91-7219-657-2.

No 627 **Vadim Engelson:** Tools for Design, Interactive Simulation, and Visualization of Object-Oriented Models in Scientific Computing, 2000, ISBN 91-7219-709-9.

No 637 **Esa Falkenroth:** Database Technology for Control and Simulation, 2000, ISBN 91-7219-766-8.

No 639 **Per-Arne Persson:** Bringing Power and Knowledge Together: Information Systems Design for Autonomy and Control in Command Work, 2000, ISBN 91-7219-796-X.

No 660 **Erik Larsson:** An Integrated System-Level Design for Testability Methodology, 2000, ISBN 91-7219-890-7.

No 688 **Marcus Bjäreland:** Model-based Execution Monitoring, 2001, ISBN 91-7373-016-5.

No 689 **Joakim Gustafsson:** Extending Temporal Action Logic, 2001, ISBN 91-7373-017-3.

No 720 **Carl-Johan Petri:** Organizational Information Provision - Managing Mandatory and Discretionary Use of Information Technology, 2001, ISBN-91-7373-126-9.

No 724 **Paul Scerri:** Designing Agents for Systems with Adjustable Autonomy, 2001, ISBN 91 7373 207 9.

No 725 **Tim Heyer:** Semantic Inspection of Software Artifacts: From Theory to Practice, 2001, ISBN 91 7373 208 7.

No 726 **Pär Carlshamre:** A Usability Perspective on Requirements Engineering - From Methodology to Product Development, 2001, ISBN 91 7373 212 5.

No 732 **Juha Takkinen:** From Information Management to Task Management in Electronic Mail, 2002, ISBN 91 7373 258 3.

No 745 **Johan Åberg:** Live Help Systems: An Approach to Intelligent Help for Web Information Systems, 2002, ISBN 91-7373-311-3.

No 746 **Rego Granlund:** Monitoring Distributed Teamwork Training, 2002, ISBN 91-7373-312-1.

No 757 **Henrik André-Jönsson:** Indexing Strategies for Time Series Data, 2002, ISBN 917373-346-6.

No 747 **Anneli Hagdahl:** Development of IT-supported Interorganisational Collaboration - A Case Study in the Swedish Public Sector, 2002, ISBN 91-7373-314-8.

No 749 **Sofie Pilemalm:** Information Technology for Non-Profit Organisations - Extended Participatory Design of an Information System for Trade Union Shop Stewards, 2002, ISBN 91-7373-318-0.

No 765 **Stefan Holmlid:** Adapting users: Towards a theory of use quality, 2002, ISBN 91-7373-397-0.

No 771 **Magnus Morin:** Multimedia Representations of Distributed Tactical Operations, 2002, ISBN 91-7373-421-7.

No 772 **Pawel Pietrzak:** A Type-Based Framework for Locating Errors in Constraint Logic Programs, 2002, ISBN 91-7373-422-5.

No 758 **Erik Berglund:** Library Communication Among Programmers Worldwide, 2002, ISBN 91-7373-349-0.

No 774 **Choong-ho Yi:** Modelling Object-Oriented Dynamic Systems Using a Logic-Based Framework, 2002, ISBN 91-7373-424-1.

No 779 **Mathias Broxvall:** A Study in the Computational Complexity of Temporal Reasoning, 2002, ISBN 91-7373-440-3.

No 793 **Asmus Pandikow:** A Generic Principle for Enabling Interoperability of Structured and Object-Oriented Analysis and Design Tools, 2002, ISBN 91-7373-479-9.

No 785 **Lars Hult:** Publika Informationstjänster. En studie av den Internetbaserade encyklopedins bruksegenskaper, 2003, ISBN 91-7373-461-6.

No 800 **Lars Taxén:** A Framework for the Coordination of Complex Systems´ Development, 2003, ISBN 91-7373-604-X

No 808 **Klas Gäre:** Tre perspektiv på förväntningar och förändringar i samband med införande av informationssystem, 2003, ISBN 91-7373-618-X.

No 821 **Mikael Kindborg:** Concurrent Comics - programming of social agents by children, 2003, ISBN 91-7373-651-1.

No 823 **Christina Ölvingson:** On Development of Information Systems with GIS Functionality in Public Health Informatics: A Requirements Engineering Approach, 2003, ISBN 91-7373-656-2.

No 828 **Tobias Ritzau:** Memory Efficient Hard Real-Time Garbage Collection, 2003, ISBN 91-7373-666-X.

No 833 **Paul Pop:** Analysis and Synthesis of Communication-Intensive Heterogeneous Real-Time Systems, 2003, ISBN 91-7373-683-X.

No 852 **Johan Moe:** Observing the Dynamic Behaviour of Large Distributed Systems to Improve Development and Testing – An Empirical Study in Software Engineering, 2003, ISBN 91-7373-779-8.

No 867 **Erik Herzog:** An Approach to Systems Engineering Tool Data Representation and Exchange, 2004, ISBN 91-7373-929-4.

No 872 **Aseel Berglund:** Augmenting the Remote Control: Studies in Complex Information Navigation for Digital TV, 2004, ISBN 91-7373-940-5.

No 869 **Jo Skåmedal:** Telecommuting's Implications on Travel and Travel Patterns, 2004, ISBN 91-7373-935-9.

No 870 **Linda Askenäs:** The Roles of IT - Studies of Organising when Implementing and Using Enterprise Systems, 2004, ISBN 91-7373-936-7.

No 874 **Annika Flycht-Eriksson:** Design and Use of Ontologies in Information-Providing Dialogue Systems, 2004, ISBN 91-7373-947-2.

No 873 **Peter Bunus:** Debugging Techniques for Equation-Based Languages, 2004, ISBN 91-7373-941-3.

No 876 **Jonas Mellin:** Resource-Predictable and Efficient Monitoring of Events, 2004, ISBN 91-7373-956-1.

No 883 **Magnus Bång:** Computing at the Speed of Paper: Ubiquitous Computing Environments for Healthcare Professionals, 2004, ISBN 91-7373-971-5

No 882 **Robert Eklund:** Disfluency in Swedish human-human and human-machine travel booking dialogues, 2004, ISBN 91-7373-966-9.

No 887 **Anders Lindström:** English and other Foreign Linguistic Elements in Spoken Swedish. Studies of Productive Processes and their Modelling using Finite-State Tools, 2004, ISBN 91-7373-981-2.

No 889 **Zhiping Wang:** Capacity-Constrained Production-inventory systems - Modelling and Analysis in both a traditional and an e-business context, 2004, ISBN 91-85295-08-6.

No 893 **Pernilla Qvarfordt:** Eyes on Multimodal Interaction, 2004, ISBN 91-85295-30-2.

No 910 **Magnus Kald:** In the Borderland between Strategy and Management Control - Theoretical Framework and Empirical Evidence, 2004, ISBN 91-85295-82-5.

No 918 **Jonas Lundberg:** Shaping Electronic News: Genre Perspectives on Interaction Design, 2004, ISBN 91-85297-14-3.

No 900 **Mattias Arvola:** Shades of use: The dynamics of interaction design for sociable use, 2004, ISBN 91-85295-42-6.

No 920 **Luis Alejandro Cortés:** Verification and Scheduling Techniques for Real-Time Embedded Systems, 2004, ISBN 91-85297-21-6.

No 929 **Diana Szentivanyi:** Performance Studies of Fault-Tolerant Middleware, 2005, ISBN 91-85297-58-5.

No 933 **Mikael Cäker:** Management Accounting as Constructing and Opposing Customer Focus: Three Case Studies on Management Accounting and Customer Relations, 2005, ISBN 91-85297-64-X.

No 937 **Jonas Kvarnström:** TALplanner and Other Extensions to Temporal Action Logic, 2005, ISBN 91-85297-75-5.

No 938 **Bourhane Kadmiry:** Fuzzy Gain-Scheduled Visual Servoing for Unmanned Helicopter, 2005, ISBN 91-85297-76-3.

No 945 **Gert Jervan:** Hybrid Built-In Self-Test and Test Generation Techniques for Digital Systems, 2005, ISBN: 91-85297-97-6.

No 946 **Anders Arpteg:** Intelligent Semi-Structured Information Extraction, 2005, ISBN 91-85297-98-4.

No 947 **Ola Angelsmark:** Constructing Algorithms for Constraint Satisfaction and Related Problems - Methods and Applications, 2005, ISBN 91-85297-99-2.

No 963 **Calin Curescu:** Utility-based Optimisation of Resource Allocation for Wireless Networks, 2005, ISBN 91-85457-07-8.

No 972 **Björn Johansson:** Joint Control in Dynamic Situations, 2005, ISBN 91-85457-31-0.

No 974 **Dan Lawesson:** An Approach to Diagnosability Analysis for Interacting Finite State Systems, 2005, ISBN 91-85457-39-6.

No 979 **Claudiu Duma:** Security and Trust Mechanisms for Groups in Distributed Services, 2005, ISBN 91-85457-54-X.

No 983 **Sorin Manolache:** Analysis and Optimisation of Real-Time Systems with Stochastic Behaviour, 2005, ISBN 91-85457-60-4.

No 986 **Yuxiao Zhao:** Standards-Based Application Integration for Business-to-Business Communications, 2005, ISBN 91-85457-66-3.

No 1004 **Patrik Haslum:** Admissible Heuristics for Automated Planning, 2006, ISBN 91-85497-28-2.

No 1005 **Aleksandra Tešanovic:** Developing Reusable and Reconfigurable Real-Time Software using Aspects and Components, 2006, ISBN 91-85497-29-0.

No 1008 **David Dinka:** Role, Identity and Work: Extending the design and development agenda, 2006, ISBN 91-85497-42-8.

No 1009 **Iakov Nakhimovski:** Contributions to the Modeling and Simulation of Mechanical Systems with Detailed Contact Analysis, 2006, ISBN 91-85497-43-X.

No 1013 **Wilhelm Dahllöf:** Exact Algorithms for Exact Satisfiability Problems, 2006, ISBN 91-85523-97-6.

No 1016 **Levon Saldamli:** PDEModelica - A High-Level Language for Modeling with Partial Differential Equations, 2006, ISBN 91-85523-84-4.

No 1017 **Daniel Karlsson:** Verification of Component-based Embedded System Designs, 2006, ISBN 91-85523-79-8

No 1018 **Ioan Chisalita:** Communication and Networking Techniques for Traffic Safety Systems, 2006, ISBN 91-85523-77-1.

No 1019 **Tarja Susi:** The Puzzle of Social Activity - The Significance of Tools in Cognition and Cooperation, 2006, ISBN 91-85523-71-2.

No 1021 **Andrzej Bednarski:** Integrated Optimal Code Generation for Digital Signal Processors, 2006, ISBN 91-85523-69-0.

No 1022 **Peter Aronsson:** Automatic Parallelization of Equation-Based Simulation Programs, 2006, ISBN 91-85523-68-2.

No 1030 **Robert Nilsson:** A Mutation-based Framework for Automated Testing of Timeliness, 2006, ISBN 91-85523-35-6.

No 1034 **Jon Edvardsson:** Techniques for Automatic Generation of Tests from Programs and Specifications, 2006, ISBN 91-85523-31-3.

No 1035 **Vaida Jakoniene:** Integration of Biological Data, 2006, ISBN 91-85523-28-3.

No 1045 **Genevieve Gorrell:** Generalized Hebbian Algorithms for Dimensionality Reduction in Natural Language Processing, 2006, ISBN 91-85643-88-2.

No 1051 **Yu-Hsing Huang:** Having a New Pair of Glasses - Applying Systemic Accident Models on Road Safety, 2006, ISBN 91-85643-64-5.

No 1054 **Åsa Hedenskog:** Perceive those things which cannot be seen - A Cognitive Systems Engineering perspective on requirements management, 2006, ISBN 91-85643-57-2.

No 1061 **Cécile Åberg:** An Evaluation Platform for Semantic Web Technology, 2007, ISBN 91-85643-31-9.

No 1073 **Mats Grindal:** Handling Combinatorial Explosion in Software Testing, 2007, ISBN 978-91-85715-74-9.

No 1075 **Almut Herzog:** Usable Security Policies for Runtime Environments, 2007, ISBN 978-91-85715-65-7.

No 1079 **Magnus Wahlström:** Algorithms, measures, and upper bounds for Satisfiability and related problems, 2007, ISBN 978-91-85715-55-8.

No 1083 **Jesper Andersson:** Dynamic Software Architectures, 2007, ISBN 978-91-85715-46-6.

No 1086 **Ulf Johansson:** Obtaining Accurate and Comprehensible Data Mining Models - An Evolutionary Approach, 2007, ISBN 978-91-85715-34-3.

No 1089 **Traian Pop:** Analysis and Optimisation of Distributed Embedded Systems with Heterogeneous Scheduling Policies, 2007, ISBN 978-91-85715-27-5.

No 1091 **Gustav Nordh:** Complexity Dichotomies for CSP-related Problems, 2007, ISBN 978-91-85715-20-6.

No 1106 **Per Ola Kristensson:** Discrete and Continuous Shape Writing for Text Entry and Control, 2007, ISBN 978-91-85831-77-7.

No 1110 **He Tan:** Aligning Biomedical Ontologies, 2007, ISBN 978-91-85831-56-2.

No 1112 **Jessica Lindblom:** Minding the body - Interacting socially through embodied action, 2007, ISBN 978-91-85831-48-7.

No 1113 **Pontus Wärnestål:** Dialogue Behavior Management in Conversational Recommender Systems, 2007, ISBN 978-91-85831-47-0.

No 1120 **Thomas Gustafsson:** Management of Real-Time Data Consistency and Transient Overloads in Embedded Systems, 2007, ISBN 978-91-85831-33-3.

No 1127 **Alexandru Andrei:** Energy Efficient and Predictable Design of Real-time Embedded Systems, 2007, ISBN 978-91-85831-06-7.

No 1139 **Per Wikberg:** Eliciting Knowledge from Experts in Modeling of Complex Systems: Managing Variation and Interactions, 2007, ISBN 978-91-85895-66-3.

No 1143 **Mehdi Amirijoo:** QoS Control of Real-Time Data Services under Uncertain Workload, 2007, ISBN 978-91-85895-49-6.

No 1150 **Sanny Syberfeldt:** Optimistic Replication with Forward Conflict Resolution in Distributed Real-Time Databases, 2007, ISBN 978-91-85895-27-4.

No 1155 **Beatrice Alenljung:** Envisioning a Future Decision Support System for Requirements Engineering - A Holistic and Human-centred Perspective, 2008, ISBN 978-91-85895-11-3.

No 1156 **Artur Wilk:** Types for XML with Application to Xcerpt, 2008, ISBN 978-91-85895-08-3.

No 1183 **Adrian Pop:** Integrated Model-Driven Development Environments for Equation-Based Object-Oriented Languages, 2008, ISBN 978-91-7393-895-2.

No 1185 **Jörgen Skågeby:** Gifting Technologies - Ethnographic Studies of End-users and Social Media Sharing, 2008, ISBN 978-91-7393-892-1.

No 1187 **Imad-Eldin Ali Abugessaisa:** Analytical tools and information-sharing methods supporting road safety organizations, 2008, ISBN 978-91-7393-887-7.

No 1204 **H. Joe Steinhauer:** A Representation Scheme for Description and Reconstruction of Object Configurations Based on Qualitative Relations, 2008, ISBN 978-91-7393-823-5.

No 1222 **Anders Larsson:** Test Optimization for Core-based System-on-Chip, 2008, ISBN 978-91-7393-768-9.

No 1238 **Andreas Borg:** Processes and Models for Capacity Requirements in Telecommunication Systems, 2009, ISBN 978-91-7393-700-9.

No 1240 **Fredrik Heintz:** DyKnow: A Stream-Based Knowledge Processing Middleware Framework, 2009, ISBN 978-91-7393-696-5.

No 1241 **Birgitta Lindström:** Testability of Dynamic Real-Time Systems, 2009, ISBN 978-91-7393-695-8.

No 1244 **Eva Blomqvist:** Semi-automatic Ontology Construction based on Patterns, 2009, ISBN 978-91-7393-683-5.

No 1249 **Rogier Woltjer:** Functional Modeling of Constraint Management in Aviation Safety and Command and Control, 2009, ISBN 978-91-7393-659-0.

No 1260 **Gianpaolo Conte:** Vision-Based Localization and Guidance for Unmanned Aerial Vehicles, 2009, ISBN 978-91-7393-603-3.

No 1262 **AnnMarie Ericsson:** Enabling Tool Support for Formal Analysis of ECA Rules, 2009, ISBN 978-91-7393-598-2.

No 1266 **Jiri Trnka:** Exploring Tactical Command and Control: A Role-Playing Simulation Approach, 2009, ISBN 978-91-7393-571-5.

No 1268 **Bahlol Rahimi:** Supporting Collaborative Work through ICT - How End-users Think of and Adopt Integrated Health Information Systems, 2009, ISBN 978-91-7393-550-0.

No 1274 **Fredrik Kuivinen:** Algorithms and Hardness Results for Some Valued CSPs, 2009, ISBN 978-91-7393-525-8.

No 1281 **Gunnar Mathiason:** Virtual Full Replication for Scalable Distributed Real-Time Databases, 2009, ISBN 978-91-7393-503-6.

No 1290 **Viacheslav Izosimov:** Scheduling and Optimization of Fault-Tolerant Distributed Embedded Systems, 2009, ISBN 978-91-7393-482-4.

No 1294 **Johan Thapper:** Aspects of a Constraint Optimisation Problem, 2010, ISBN 978-91-7393-464-0.

No 1306 **Susanna Nilsson:** Augmentation in the Wild: User Centered Development and Evaluation of Augmented Reality Applications, 2010, ISBN 978-91-7393-416-9.

No 1313 **Christer Thörn:** On the Quality of Feature Models, 2010, ISBN 978-91-7393-394-0.

No 1321 **Zhiyuan He:** Temperature Aware and Defect-Probability Driven Test Scheduling for System-on-Chip, 2010, ISBN 978-91-7393-378-0.

No 1333 **David Broman:** Meta-Languages and Semantics for Equation-Based Modeling and Simulation, 2010, ISBN 978-91-7393-335-3.

No 1337 **Alexander Siemers:** Contributions to Modelling and Visualisation of Multibody Systems Simulations with Detailed Contact Analysis, 2010, ISBN 978-91-7393-317-9.

No 1354 **Mikael Asplund:** Disconnected Discoveries: Availability Studies in Partitioned Networks, 2010, ISBN 978-91-7393-278-3.

No 1359 **Jana Rambusch**: Mind Games Extended: Understanding Gameplay as Situated Activity, 2010, ISBN 978-91-7393-252-3.

No 1373 **Sonia Sangari**: Head Movement Correlates to Focus Assignment in Swedish,2011,ISBN 978-91-7393-154-0.

No 1374 **Jan-Erik Källhammer**: Using False Alarms when Developing Automotive Active Safety Systems, 2011, ISBN 978-91-7393-153-3.

No 1375 **Mattias Eriksson**: Integrated Code Generation, 2011, ISBN 978-91-7393-147-2.

No 1381 **Ola Leifler**: Affordances and Constraints of Intelligent Decision Support for Military Command and Control – Three Case Studies of Support Systems, 2011, ISBN 978-91-7393-133-5.

No 1386 **Soheil Samii**: Quality-Driven Synthesis and Optimization of Embedded Control Systems, 2011, ISBN 978-91-7393-102-1.

No 1419 **Erik Kuiper**: Geographic Routing in Intermittently-connected Mobile Ad Hoc Networks: Algorithms and Performance Models, 2012, ISBN 978-91-7519-981-8.

No 1451 **Sara Stymne**: Text Harmonization Strategies for Phrase-Based Statistical Machine Translation, 2012, ISBN 978-91-7519-887-3.

Linköping Studies in Arts and Science

No 504 **Ing-Marie Jonsson:** Social and Emotional Characteristics of Speech-based In-Vehicle Information Systems: Impact on Attitude and Driving Behaviour, 2009, ISBN 978-91-7393-478-7.

Linköping Studies in Statistics

No 9 **Davood Shahsavani:** Computer Experiments Designed to Explore and Approximate Complex Deterministic Models, 2008, ISBN 978-91-7393-976-8.

No 10 **Karl Wahlin:** Roadmap for Trend Detection and Assessment of Data Quality, 2008, ISBN 978-91-7393-792-4.

No 11 **Oleg Sysoev:** Monotonic regression for large multivariate datasets, 2010, ISBN 978-91-7393-412-1.

No 13 **Agné Burauskaite-Harju:** Characterizing Temporal Change and Inter-Site Correlations in Daily and Sub-daily Precipitation Extremes, 2011, ISBN 978-91-7393-110-6.

Linköping Studies in Information Science

No 1 **Karin Axelsson:** Metodisk systemstrukturering- att skapa samstämmighet mellan informationssystemarkitektur och verksamhet, 1998. ISBN-9172-19-296-8.

No 2 **Stefan Cronholm:** Metodverktyg och användbarhet - en studie av datorstödd metodbaserad systemutveckling, 1998, ISBN-9172-19-299-2.

No 3 **Anders Avdic:** Användare och utvecklare - om anveckling med kalkylprogram, 1999. ISBN-91-7219-606-8.

No 4 **Owen Eriksson:** Kommunikationskvalitet hos informationssystem och affärsprocesser, 2000, ISBN 91-7219-811-7.

No 5 **Mikael Lind:** Från system till process - kriterier för processbestämning vid verksamhetsanalys, 2001, ISBN 91-7373-067-X.

No 6 **Ulf Melin:** Koordination och informationssystem i företag och nätverk, 2002, ISBN 91-7373-278-8.

No 7 **Pär J. Ågerfalk:** Information Systems Actability - Understanding Information Technology as a Tool for Business Action and Communication, 2003, ISBN 91-7373-628-7.

No 8 **Ulf Seigerroth:** Att förstå och förändra systemutvecklingsverksamheter - en taxonomi för metautveckling, 2003, ISBN91-7373-736-4.

No 9 **Karin Hedström:** Spår av datoriseringens värden – Effekter av IT i äldreomsorg, 2004, ISBN 91-7373-963-4.

No 10 **Ewa Braf:** Knowledge Demanded for Action - Studies on Knowledge Mediation in Organisations, 2004, ISBN 91-85295-47-7.

No 11 **Fredrik Karlsson:** Method Configuration method and computerized tool support, 2005, ISBN 91-85297-48-8.

No 12 **Malin Nordström:** Styrbar systemförvaltning - Att organisera systemförvaltningsverksamhet med hjälp av effektiva förvaltningsobjekt, 2005, ISBN 91-85297-60-7.

No 13 **Stefan Holgersson:** Yrke: POLIS - Yrkeskunskap, motivation, IT-system och andra förutsättningar för polisarbete, 2005, ISBN 91-85299-43-X.

No 14 **Benneth Christiansson, Marie-Therese Christiansson:** Mötet mellan process och komponent - mot ett ramverk för en verksamhetsnära kravspecifikation vid anskaffning av komponentbaserade informationssystem, 2006, ISBN 91-85643-22-X.