Machine translation: statistical approach with additional linguistic knowledge

Von der Fakultät für Mathematik, Informatik
und Naturwissenschaften der
RWTH Aachen University
zur Erlangung des akademischen Grades einer
Doktorin der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Diplom–Ingenieurin

Maja Popović

aus Belgrad, Serbien

Berichter: Universitätsprofessor Dr.-Ing. Hermann Ney
Professor Dr. Andy Way

Tag der mündlichen Prüfung: 30. April 2009

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

Acknowledgments

I would like to express my gratitude to all the people who supported and accompanied me during the preparation of this work.

First, I would like to express my gratitude to my advisor Professor Dr.-Ing. Hermann Ney, head of the Lehrstuhl für Informatik 6 at the RWTH Aachen University. This thesis would not have been possible without his advices and patience. I am very grateful that he gave me the possibility to attend various conferences, workshops and meetings.

I would also like to thank Professor Dr. Andy Way from the School of Computing at the Dublin City University for agreeing to review this thesis, for his useful comments and suggestions, and for his interest in this work.

I am also very grateful to Professor Dr.-Ing. Slobodan Jovičić from the Faculty of Electrical Engineering at the University of Belgrade for all suggestions and advices.

Many thanks to all the people at the Lehrstuhl für Informatik 6 for the great working atmosphere. Also many thanks for the great "non-working" atmosphere in "Café Bender", in the department excursions, the Christmas dinners, the "good movie" sessions as well the "bad movie" sessions, the strange Power-Point sessions, and of course the Card&Board Game sessions. Furthermore, I would like to thank the secretaries and the system administrators for their continuous support. Special thanks to all those who helped me in writing this thesis by proofreading it and adding missing articles. And special thanks to Nicola Ueffing, Franz Josef Och and Ralf Schlüter for support and help at my beginnings in Aachen and at i6.

I would like to say "thanks a lot" to the people who worked with me in the field of morpho-syntactic information and error analysis in the framework of the TC-STAR project: Adriá de Gispert, Patrik Lambert and Deepa Gupta, as well as Rafael Banchs and Marcello Federico. Many thanks to Necip Fazil Ayan for providing results of human error analysis for GALE texts.

To David, mi ludi zli informatičar: ¡Muchas gracias para todo! (incluido leer mi tesis en la "1, 2, 3" y aguantar mi agresividad en las ultimas semanas ;-)

And to my mother Biljana and my sister Nikica: Znate sve! :-)

This thesis is based on work carried out during my time as a research scientist at the Lehrstuhl für Informatik 6 at the RWTH Aachen University, Germany. The work was partially funded by European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistical Methods for Written Language Translation" (Ne572/5) and by the Defense Advanced Research Project Agency (DARPA) under contract No. HR0011-06-C-0023 (GALE).

Abstract

In this thesis, three possible aspects of using linguistic (i.e. morpho-syntactic) knowledge for statistical machine translation are described: the treatment of syntactic differences between source and target language using source POS tags, statistical machine translation with a small amount of bilingual training data, and automatic error analysis of translation output.

Reorderings in the source language based on the POS tags are systematically investigated: local reorderings of nouns and adjectives for the Spanish–English language pair and long-range reorderings of verbs for the German–English language pair. Both types of reorderings result in better performance of the translation system, local reordering being more important for the scarce training corpora.

For such corpora, strategies for achieving an acceptable translation quality by applying appropriate morpho-syntactic transformations are exploited for three language pairs: Spanish–English, German–English and Serbian–English. Very scarce task-specific corpora as well as conventional dictionaries are used as bilingual training material. In addition to conventional dictionaries, the use of phrasal lexica is proposed and investigated.

A framework for automatic analysis and classification of actual errors in translation output based on combining existing automatic evaluation measures with linguistic information is presented. Experiments on different types of corpora and various language pairs show that the results of automatic error analysis correlate very well with the results of human evaluation. The new metrics based on analysed error categories are used for comparison of different translation systems trained on various sizes of texts with and without morpho-syntactic transformations.

For improving the quality of a statistical machine translation system by the use of morpho-syntactic information, the choice of the method and the significance of improvements strongly depend on the language pair, the translation direction and the nature of the corpus. Error analysis of the translation output is important in order to define weak points of the system and apply methods for improvement in the optimal way.

Zusammenfassung

In dieser Arbeit werden drei Aspekte der Verwendung linguistischen (morpho-syntaktischen) Wissens in der statistischen Übersetzung dargestellt: Behandlung der syntaktischen Unterschiede zwischen Quellsprache und Zielsprache unter Zuhilfenahme von POS-Informationen, statistische Übersetzung bei geringen Mengen an Trainingdaten und automatische Fehleranalyse von Übersetzungsergebnissen.

Umordnungen in der Quellsprache basierend auf POS-Information werden systematisch untersucht: lokale Umordnungen von Nomen und Adjektiven für das Sprachpaar Spanisch-Englisch sowie weiträumige Umordnungen von Verben für das Sprachpaar Deutsch-Englisch. Beide Typen von Umordnungen führen zu verbesserter Übersetzungsqualität; die lokalen Umordnungen stellen sich als besonderes hilfreich für die Übersetzung bei geringen Mengen an bilingualen Trainingdaten heraus.

Für solche Übersetzungsysteme, wo nur geringe Mengen bilingualer Trainingsdaten verfügbar sind, werden morpho-syntaktische Transformationen auf ihre Eignung untersucht, um eine akzeptable Übersetzungsqualität zu erreichen. Systematische Experimenten werden auf drei verschieden Sprachpaaren durchgeführt: Spanisch-Englisch, Deutsch-Englisch und Serbisch-Englisch. Sehr kleinvolumige aufgabebezogene Daten, sowie konventionelle Wörterbücher, werden als bilinguales Trainingsmaterial benutzt. Neben den Wörterbücher werden auch phrasale Lexika vorgeschlagen und untersucht.

Es wird ein Rahmenwerk für die automatische Analyse und Klassifizierung von Fehlern basierend auf verbreiteten Fehlermassen und auf linguistischem Wissen vorgestellt. Experimente auf verschiedene Korpora und Sprachpaaren zeigen, dass die Ergebnisse der automatischen Fehleranalyse eine hohe Korelation mit den Ergebnissen menschlicher Fehleranalyse aufweisen. Die neu eingeführten auf den analysierten Fehlerkategorien beruhenden Fehlerraten werden für einen Vergleich verschiedener Übersetzungsysteme benutzt. Diese Systemen wurden zuvor auf unterschiedlichen bilingualen Datenmengen trainiert, sowohl mit als auch ohne Verwendung morpho-syntaktischer Transformationen.

Die Wahl der Methoden der Verwendung linguistischen Wissen zur Verbesserung eines statistischen Übersetzungsystems hängt ebenso wie die Signifikanz der dadurch erreichten Verbesserungen sehr vom zugrundeliegenden Sprachpaar, der Übersetzungsrichtung und der Art des Korpus ab. Fehleranalyse erweist sich als wichtig, um die Schwächen eines Übersetzungssystems zu entdecken und geeignete Methoden für eine optimale Verbesserung zu entwickeln.

Contents

List of Tables

List of Figures

1 Introduction

The statistical approach to machine translation has received a growing interest over the last decade, and different concepts and algorithms have been investigated. The use of linguistic (usually morpho-syntactic) information can improve the quality of a statistical machine translation (SMT) system. The significance of the obtained improvements depends on the language pair, the translation direction and the nature of the corpus. An error analysis of the generated output is important in order to choose appropriate methods, i.e. identification of the main error sources and possibilities to overcome these problems.

The goal of this work is to systematically investigate three aspects of the use of linguistic information for statistical machine translation:

- "harmonising" syntactic (word order) differences between two languages using part-of-speech (POS) information;
- building SMT systems with scarce bilingual training data;
- automatic analysis and classification of errors in the translation output.

The languages investigated in this thesis are Spanish, German, Serbian and English. The proposed methods are language group-specific: for example, the methods appropriate for the tasks involving the Spanish language can be extended to other Roman languages like French, Italian, etc., and the methods for Serbian could be extended to other Slavic languages like Czech, Slovenian, etc. Treatment of German verbs could be succesfully applied to Dutch, and splitting compound words can be extended to other languages with compositional morphology, such as Dutch and Finnish.

1.1 Statistical machine translation and linguistic knowledge

The goal of machine translation is the automatic translation of a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$ into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. The statistical approach to machine translation is defined as a decision problem: given a source sequence f_1^J, the optimal translation is the target sequence e_1^I which maximises the posterior probability $Pr(e_1^I|f_1^J)$. According to Bayes' rule, this posterior probability can be divided into two probabilities: the target language model probability $Pr(e_1^I)$ and the translation model probability $Pr(f_1^J|e_1^I)$. The translation model probability describes the correspondences between the words in the source and the target sequence. These correspondences are described by alignments which assign target word positions to each source word position. The language model probability describes the word order of the target language. These two probabilities can be modelled independently of each other.

Many state-of-the-art SMT systems are based on an alternative approach: the posterior probability is directly modelled as a log-linear combination of different models [Och & Ney 02, Och & Ney 04]. This approach allows the integration of many different sub-models whose weights can be directly optimised with respect to some evaluation criterion. The noisy channel model described previously can be interpreted as a special case of the log-linear model. Most state-of-the-art systems are based on phrase pairs [Zens & Och$^+$ 02]. These phrases are simply sequences of words and not necessarily phrases in the linguistic sense.

The main advantage of the statistical approach to machine translation is that it is able to learn from data automatically, and therefore it can easily be adapted to new domains and language pairs. However, additional knowledge sources can help to increase the quality of the translation. In this thesis we investigate three possibilities to use additional linguistic knowledge within the statistical framework.

1.1.1 POS-based word reorderings

The difference in word order between languages is one of the main difficulties a machine translation (MT) system has to deal with. These can range from small differences in word order (as in the case between most European languages) to a completely different sentence structure (e.g. in the case of translation from Chinese into English).

Two basic approaches are used to overcome the reordering problem for the phrase-based approach to SMT. One is to allow reorderings at the word level and then allow the system to decide which word order fits best with the translation model. Normally this is implemented by building a word graph and then allowing a "monotone" translation along this word graph [Zens & Och$^+$ 02]. Because allowing all the possible reorderings increases the complexity of the machine translation process exponentially (and in fact makes the translation problem NP-complete [Knight 99]), usually so-called reordering constraints are used. The most widely used are the IBM [Berger & Brown$^+$ 96] and the local reordering constraints [Kanthak & Vilar$^+$ 05] which are more succesful for languages with a similar word order.

Another possibility is to allow the phrase-based system to skip some parts of the source sentence and delay its translation to a later point in the search process. How many words to skip at each step and which costs to assign to such reorderings are usually dependent on the phrase model. Thus, although it would be possible to represent these reorderings in form of a graph, as with the previous reorderings, the process is normally integrated directly in the search procedure. This approach allows for longer range reorderings and is therefore normally used for languages with larger differences in sentence structure.

Both of these approaches, however, have a common problem, namely they allow for "too many" reorderings, i.e. they have relatively loose restrictions. This has two immediate consequences: on the one hand, the large number of permitted reorderings increases the complexity of the search process. On the other hand, the greater liberty in choosing the word order to

translate can accentuate the modelling errors of the translation system used.[1] Therefore, the reordering constraints and the reordering parameters must be carefully tuned.

Additional linguistic knowledge can help to overcome these difficulties. If the differences between the word order of the involved languages are known, we can instruct the system how to deal with them. An example of translation from Spanish into English can help to clarify this point. In the Spanish language most adjectives (although not all) are placed after the corresponding noun, whereas in English the opposite is true. We can, therefore, define rules for moving the adjectives in the Spanish source sentence to preceed the corresponding noun before the translation process. In this thesis we explore how to exploit linguistic knowledge about the word order for improving the MT quality using POS information. We focus on two European language pairs, Spanish–English, where short-range (local) word reorderings are needed, and German–English, where long-range word reorderings can be found.

1.1.2 Translation with scarce bilingual resources

One of the main characteristics of the statistical approach to machine translation is its ability to automatically learn from training examples. In order to build a system for a language pair just a set of bilingual parallel sentences is needed. However, while being one of the main advantages of this approach, it is at the same time one of its drawbacks. The translation quality largely depends on the size of the available training data. Although for many language pairs a high amount of parallel data is available,[2] for the vast majority of language pairs only small amounts of data is available. This is especially true for minority languages or languages in danger of extinction. Furthermore, the production of high-quality bilingual corpora is still a time-consuming and expensive task.

In this thesis we will investigate systematically the impact of corpus size on translation quality for three European language pairs. We will also consider how to effectively use standard ("word-based") dictionaries as the only bilingual data available or to increase the quality of a translation system when only a very limited amount of bilingual text is available.

Additionally, one of the main problems of the statistical approach is that it is only able to learn translations of words as whole units, i.e. no generalisation for different forms of the same base word can be learnt with state-of-the-art approaches. For example, if the system has only seen a limited amount of verb forms (for richly inflected languages), it will not be able to generalise to new verb forms not seen in the training data. This limitation is especially important in the case of scarce training resources.

In this work we will show how to use additional linguistic knowledge about the morphology of words to overcome this limitation. This approach will be of great advantage for situations when only a very limited amount of training data is available.

[1] It should not be forgotten that we are dealing with *models* of the probability distributions, not with the "true" probability distributions themselves. As such, many modelling assumptions are made, most of them highly approximative or even incorrect.

[2] For example between some European languages in the form of the proceedings of the European parliament, or for the translation from Chinese or Arabic into English, due to greater attention on these pairs in recent projects such as TC-STAR, GALE, etc.

1.1.3 Automatic error analysis of translation output

The evaluation of MT output is an important task, yet at the same time a difficult problem for progress on the field to be made. Because there is no unique reference translation for a text (like for example in speech recognition), automatic measures are hard to define. Human evaluation, while of course providing (at least in principle) the most reliable judgements, is costly and time-consuming. A great deal of effort has been spent on finding measures that correlate well with human judgements when determining which one of a set of translation systems is the best (be it different versions of the same system in the development phase or a set of "competing" systems as, for example, in a machine translation evaluation).

However, most of the work has been focused just on "best-worst" decisions, i.e. finding a ranking between different machine translation systems. Although being useful information and helping in the continuous improvement of machine translation systems, MT researchers often would find it helpful to have additional information about their systems. What are the strengths of their systems? Where do they make errors? Does a particular modification improve some aspect of the system, although perhaps it does not improve the overall score in terms of one of the standard measures? Hardly any systematic work has been done in this direction and developers must resort to looking into the translation outputs in order to obtain an insight into the actual problems of their systems.

In this thesis we propose a framework for automatic error analysis of MT output. We extend the standard evaluation measures by the use of linguistic knowledge to show which kind of errors an MT system produces and present methods in order to find the problematic sections of the produced translations.

1.2 Related work

1.2.1 Morphological and syntactic transformations for SMT

There are many publications dealing with various types of morphological and syntactic analysis, such as POS-tagging [Ratnaparkhi 96, Brants 00, Toutanova & Manning 00], discovering of morphemes [Goldsmith 01, Creutz & Lagus 02], etc. Morpho-syntactic analysis has also been used to improve the quality of speech recognition systems, for example treatment of German compound words in [Larson & Willett$^+$ 00, Larson 01] and splitting Finnish words into morphemes in [Siivola & Hirsimäki$^+$ 03].

Using this type of information in SMT systems was proposed already in the beginning by [Brown & Cocke$^+$ 90, Brown & Della Pietra$^+$ 92] for the French–English language pair, but applied and tested about one decade later on the German–English language pair [Nießen & Ney 00, Nießen 02]. In the last five years there have been a number of publications dealing with morphological and syntactic analysis and its applications to SMT which will be described in the next subsections.

Improving alignment models by different types of morpho-syntactic knowledge has been also investigated in recent years, e.g. [Toutanova & Ilhan$^+$ 02, de Gispert & Gupta$^+$ 06]. However, the relation between the improvement of alignment error rate (AER) and the translation quality is still not clear (see for example [Fraser & Marcu 06, Vilar & Popović$^+$ 06]).

1.2.1.1 Morphological transformations

The first work which reports improvements in the performance of an SMT system using morphological transformations is [Nießen & Ney 00]. They propose various methods for the German–English language pair: splitting German compound words, disambiguation of ambiguous words with POS tags, merging multi-word phrases and treatment of unseen word forms. Further work on the same language pair [Nießen & Ney 01b] introduces a hierarchical lexicon model for the translation from German into English; the German part of the corpus is enriched with the corresponding base forms and sequences of relevant POS tags. Splitting of German compounds is also dealt with in [Koehn & Knight 03]. Contrary to [Nießen & Ney 00], they do not use any morphological analyser, but propose several corpus-based methods.

The problem of rich inflectional morphology of Spanish verbs is adressed in [Ueffing & Ney 03]. They merge English words which correspond to one Spanish verb with the use of English POS tags. Inflections of Spanish verbs are also treated in [de Gispert & Mariño$^+$ 05]. They propose classification and generalisation of Spanish and English verbs based on POS tags and base forms, where each verb form or verb group is replaced by the base form of the main verb.

Morphological analysis of the Arabic language is investigated in [Lee 04]. Word segmentation into prefix, stem and suffix is applied for translation into English. [Goldwater & McClosky 05] addresses rich inflectional morphology of the Czech language for translation into English.

Although morphological transformations are not in the focus of this thesis, some transformations will be investigated for cases of scarce training corpora such as optimal use of base forms.

1.2.1.2 POS-based word reorderings

Many publications deal with the word reordering problem, but only few of them make use of linguistic knowledge about the sentence structure. The well known IBM reordering constraints for search initially proposed for the word level [Berger & Brown$^+$ 96] are based on a coverage vector which marks already translated source positions; at each position, only the first k still uncovered positions can be translated. The ITG constraints [Wu 97] are inspired by bilingual bracketing; the sentence is reordered by combining word segments, and at each step two adjacent segments are merged either in the original or in inverted order. Two reordering techniques for search, inverse IBM constraints and local reorderings within window of k positions, are proposed in [Kanthak & Vilar$^+$ 05]. Additionally, they also introduce reordering in training based on monotonisation of the word alignment. Still, they do not use any linguistic knowledge.

Using morpho-syntactic information for local word reordering transformations for the French–English language pair was suggested already in [Brown & Della Pietra$^+$ 92]. Unfortunately, they did not report any experimental results considering the effect of the reordering on the translation quality. The first application of morpho-syntactic information for word reordering in SMT is reported in [Nießen & Ney 01a] for the German–English pair. Two reordering transformations are proposed: prepending German verb prefixes to the main verb and inversion of interrogative sentences using syntactic information. Another method for harmonising word

order between the same language pair has been proposed in [Collins & Koehn+ 05]. They use a German parse tree for moving German verb prefixes, infinitives, negative particles and finite verbs towards the beginning of the clause. A similar method is applied in [Wang & Collins+ 07] on the Chinese–English language pair but for reordering of phrases instead of words. All these publications apply reordering transformations as a preprocessing step.

In the last two years several publications addressed the problem of local reorderings for the Spanish–English language pair. In [Lee & Ge 06] reordering rules are acquired from a word-aligned parallel corpus using POS tags of the source part and then applied as a preprocessing step. A similar method for extracting local reordering patterns for both translation directions is explored in [Crego & de Gispert+ 06] and [Crego & Mariño 06]. The obtained patterns are then used for the creation of word graphs which contain all possible paths. A similar approach for the Chinese–English language pair is presented in [Zhang & Zens+ 07], but shallow parsing chunks for phrase reordering are used instead of POS tags for word reordering. Extracting rules from word alignments and source language POS tags is also presented in [Rottmann & Vogel 07] for the Spanish–English and German–English language pair. These rules are then used for the creation of word graphs, but the graphs are extended with the word or POS tag context in which a reordering pattern is seen in the training data.

Some more publications have dealt with this issue: statistical machine reordering for the Spanish–English language pair where the reordering rules are extracted from word alignments along with automatically learnt word classes is proposed in [Costa-jussà & Fonollosa 06]. For the same language pair, [Kirchhoff & Yang+ 06] propose reordering of nouns and adjectives as a postprocessing step for the English output.

This thesis investigates POS-based word reorderings of the source language for two language pairs and four translation directions: Spanish–English and German–English. A word alignment of the corpus is not needed, the only necessary additional source being POS information. A parse tree or some other type of detailed information about syntax is not necessary. For the German–English language pair two novel reorderings are introduced which are important, but have not been tested in previous work, namely reordering of past participles and specific infinitive groups. In addition, reorderings for translation into German are proposed. Both variants for applying reorderings on the test corpus are investigated: rule-based reordering before training and translation as well as creation of a graph containing all possible paths with constraints.

1.2.2 Translation with scarce bilingual resources

Strategies for exploiting limited amounts of bilingual data are receiving more and more attention. In the last five years various publications have dealt with the issue of sparse bilingual corpora.

The use of conventional dictionaries to augment or replace parallel corpora has already been examined by [Brown & Della Pietra+ 93]. In [Nießen & Ney 04] and [Vogel & Monson 04] conventional dictionaries were augmented with additional morphological variations.

[Al-Onaizan & Germann+ 00] report an experiment of Tetun–English translation with a small parallel corpus, although this work is not focused on the statistical approach. The translation experiment is done by different groups including one using SMT. They found that the

human mind is very capable of deriving dependencies such as morphology, cognates, proper names, etc. and that this capability is the crucial reason for the better results produced by humans compared to corpus-based machine translation. If a program sees a particular word or phrase one thousand times during the training, it is more likely to learn a correct translation than if it sees it ten times, or never. Because of this, statistical translation techniques are less likely to work well when only a small amount of data is given.

[Callison-Burch & Osborne 03] propose a co-training method for SMT using the multilingual European Parliament corpus. Multiple translation models trained on different language pairs are used to produce new sentence pairs. They are then added to the original corpus and all translation models are retrained. The best improvements were achieved after two or three training rounds.

In [Nießen & Ney 04] the impact of the size of training corpus for SMT from German into English is investigated, and the use of a conventional dictionary and morpho-syntactic information for improving the performance is proposed. They use several types of word reorderings as well as a hierarchical lexicon based on POS tags and base forms of the German language. They report results for training on the full corpus of about sixty thousand sentences, on a small part of the corpus containing five thousand sentences and on the conventional dictionary only. Morpho-syntactic information yields significant improvements in all cases and an acceptable translation quality is also obtained with the very small corpus.

This thesis will systematically investigate various tasks involving three distinct language pairs and appropriate morpho-syntactic information. Besides the use of conventional dictionaries, the use of a phrasal lexicon as an additional knowledge source will be explored.

1.2.3 Automatic error analysis of translation output

A variety of automatic evaluation measures have been proposed and studied over the last years, some of which have been shown to be very useful tools for comparing different systems as well as for evaluating improvements within one system. The most widely used are Word Error Rate (WER), Position-independent word Error Rate (PER), the BLEU score [Papineni & Roukos$^+$ 02] and the NIST score [Doddington 02]. A General Text Matcher (GMT) approach for measuring similarity between texts based on precision, recall and F-measure is proposed and described in [Melamed & Green$^+$ 03, Turian & Shen$^+$ 03]. Recently, the Translation Edit Rate (TER) [Snover & Dorr$^+$ 06] is receiving more and more attention. It is based on the edit distance (WER) but with an additional cost for shifts of word sequences. [Leusch & Ueffing$^+$ 05] investigate preprocessing and normalisation methods for improving the evaluation using the standard measures WER, PER, BLEU and NIST. The same set of measures is examined in [Matusov & Leusch$^+$ 05] in combination with automatic sentence segmentation in order to enable evaluation of translation output without sentence boundaries (e.g. translation of speech recognition output). An extended version of BLEU which uses *n*-grams weighted according to their frequency estimated from a monolingual corpus is proposed in [Babych & Hartley 04]. The automatic metric METEOR [Banerjee & Lavie 05] uses stems and synonyms of the words. This measure counts the number of exact word matches between the output and the reference. In a second step, unmatched words are converted into stems or synonyms and then matched. The CDER mea-

sure [Leusch & Ueffing+ 06] is based on edit distance, such as the well-known WER, but allows reordering of blocks. Evaluation based on sentence structure instead of strings is proposed in [Liu & Gildea 05] and [Owczarzak & van Genabith+ 07] – the first approach uses syntactic features for evaluation, and the other is based on the dependency structure of the sentences. IQMT [Giménez & Amigó 06] is a framework for automatic evaluation in which evaluation metrics can be combined. Nevertheless, none of these measures or extensions takes into account linguistic knowledge about actual translation errors, for example what is the contribution of verbs in the overall error rate, how many full forms are wrong whereas their base forms are correct, etc. [Vilar & Xu+ 06] proposed a framework for human error analysis and error classification based on the method presented in [Llitjós & Carbonell+ 05], and a detailed analysis of the obtained results has been carried out. However, human error analysis, like any human evaluation, is a time-consuming task.

Whereas the use of linguistic knowledge for improving the performance of an SMT system is investigated in many publications for various language pairs, its use for the analysis of translation errors is still a rather unexplored area.

In this thesis, a framework for automatic error analysis based on the two standard error rates WER and PER in combination with linguistic knowledge is defined, and detailed and systematic research is performed on various tasks and language pairs.

2 Scientific goals

As discussed in the previous chapter, the main objectives of this thesis are the following aspects of using morphological and syntactic information for SMT of different language pairs:

- treatment of syntactic differences between languages using POS tags,
- systematic investigation of the trade-off between the size of the bilingual training corpus and the translation quality,
- automatic error analysis of the translation output.

Those aspects, as well as the applied methods described in this thesis are not at all independent. Rather, they intersect and complement each other.

Word reorderings based on POS tags

Although statistical alignment models capture the differences in word order between two languages, and non-monotonic search strategies are able to handle these differences in the translation process, word order is still one of the main sources of errors in SMT. Therefore it is promising to examine transformations which aim at "harmonising" the word order in corresponding sentences on the basis of some linguistic knowledge about the sentence structure. The most used approach is the extraction of reordering rules using Viterbi word alignment in combination with POS tag or chunk sequences [Costa-jussà & Fonollosa 06, Crego & Mariño 06, Rottmann & Vogel 07, Zhang & Zens$^+$ 07]. The rules are then used on the test corpus to create word graphs. Some publications propose syntactic analysis for reordering as a preprocessing step for training and translation [Nießen & Ney 01a, Collins & Koehn$^+$ 05, Wang & Collins$^+$ 07]. None of these publications investigates word graphs.

This thesis will investigate the word reorderings based only on POS information in the source language. A word alignment between the source and target corpora is not needed, nor is a deep syntactic analysis of either language. Reordering rules are applied on the test set in both ways: fixed reordering as a preprocessing step, and creation of a word graph containing the original path and all possible paths produced by reorderings. The work is focused on two language pairs and two types of reorderings: local reorderings of nouns and adjectives for the Spanish–English language pair, and long–range reorderings of verbs for the German–English language pair.

Translation with scarce bilingual resources

One of the objectives of this thesis is to achieve an acceptable translation quality with very scarce amounts of bilingual training data. Acquisition of a large bilingual parallel text for the desired domain and language pair is a costly and time–consuming task. For some language

pairs this is very hard or almost impossible. Therefore machine translation with small amounts of bilingual data are receiving more and more attention in the literature. However, the use of linguistic information as an additional knowledge source for a statistical machine translation system is hardly investigated. In [Nießen & Ney 04] several methods for translation from the German language into English are presented.

In this work, translation with a small amount of bilingual data for three distinct language pairs and different domains will be systematically investigated. Translation with small task-specific corpora will be examined, as well as translation with a conventional dictionary. Conventional dictionaries will be used in two ways: as the only available bilingual corpus, and as additional training material for a small task-specific corpus. In addition to the use of conventional dictionaries, the use of phrasal lexica will be examined. For each language pair, appropriate morpho-syntactic transformations including POS-based word reorderings will be examined in order to improve the translation perfomance.

Automatic error analysis of translation output

The main drawback of the widely-used standard automatic evaluation measures is the lack of information about the nature of actual translation errors. On the other hand, human error analysis and classification is, as human evaluation, a costly and time–consuming task. Therefore it is promising to investigate methods for the automatic analysis and classification of errors in translation output. To the best of our knowledge, this area has not been addressed yet in the literature.

One of the goals of this thesis is to set up a framework for analysis of translation errors based on automatic evaluation measures. The results of the proposed automatic error analysis will be compared with the results of human error analysis. New metrics related to particular types of translation errors will be defined in order to compare different translation systems, i.e. to see how the new metrics reflect the effects of the applied morpho-syntactic transformations as well as of different sizes of the training corpora.

3 Morpho-syntactic information

Morphology is a subdiscipline of linguistics which studies the structure of words, whereas syntax is primarily concerned with the structure of sentences, i.e. the word order and agreement in the relationship between words. The part of morphology that covers the relationship between syntax and morphology is called morphosyntax. Morpho-syntactic information can be used as additional knowledge in SMT systems in order to overcome the problems caused by the morphological and syntactic differences between two languages. Apart from this, morpho-syntactic knowledge in combination with automatic error measures can give a better overview of the nature of actual translation errors.

Morphological differences: When translating a more inflected language into English, one of the problems is a low coverage of the probabilistic lexicon. Since existing SMT systems usually regard only full forms of words, the translation of full forms which were not seen in the training corpus is not possible even if the base form or stem of the word was seen. Another problem is that an English word might correspond to only a part of a word in the other language. For example, the Spanish form of the verb “estar” (to be) in the first person plural present tense “estamos” corresponds to the two English words “we are” (the stem “esta” corresponds to the word “are” and the suffix “mos” to the word “we”. A similar problem is compositional morphology in the German language. For example, the German word “Menschenrechte” corresponds to the two English words “human rights”. Although phrase-based systems are able to handle these phenomena well, there are still a number of translation errors caused by morphological differences between the languages. Translation from English into a more inflected language is even harder because it is very difficult to choose the correct inflection. Translation between two inflected language poses even more morphological problems. Therefore the knowledge obtained from morphological analysis can help an SMT system, especially if only a small amount of bilingual training data is available.

Syntactic differences: The word order of a source language normally differs from the word order of a target language, and for some language pairs these differences are substantial. For example, in English an adjective always precedes its corresponding noun whereas in Roman languages such as Spanish, Italian and French it is usually the other way round. In the German language, the verb is often at the end of the clause, which is not the case for the majority of other languages. Differences between Chinese and English are even larger; there are long distance differences between the positions of whole phrases. In spite of the high quality of the state-of-the art phrase-based translation systems, these differences still pose difficulties. Therefore the use of syntactic information can be used as an additional knowledge source for improving the translation quality.

Error analysis: In order to obtain more information about the nature of actual translation errors, morpho-syntactic information can be used in combination with automatic evaluation measures. Syntactic differences and error analysis are the focus of this work. Some morphological transformations are also investigated, but predominantly on the small training corpora.

3.1 Basic concepts

This section presents the basic concepts regarding the morpho-syntactic knowledge used in this work. In morphology, three different principles of word formation can be distinguished:inflection, derivation and composition. Inflection is a change of the form of a word (usually by adding a suffix) to indicate a change in its grammatical function such as case, number, gender, person, tense, mood or voice. Familiar examples of inflections are the conjugations of verbs and the declensions of nouns. Inflected forms of the words do not generally appear in dictionaries. Derivation is the combination of a word with an affix, for example clear*ly*, *un*clear. Words with derivational affixes often do appear in dictionaries. Composition is the construction of a new word by joining two or more words, such as "Menschenrechte" in German. This process can lead to generally infinite vocabularies in languages like German or Finnish; new words can be generated by joining an arbitrary number of existing words.

These morphological processes are productive in verbs, nouns, adjectives, adverbs, pronouns and determiners which results in continuous fluctuation in these word classes. On the other hand, conjunctions and prepositions are not subject to the morphological processes.

The base form (also called "lemma" is the uninflected form of a word which typically serves as a primary key for a dictionary. For nouns, this is the singular nominative form, for verbs the infinitive form, for adjectives the indefinite adverbial or the singular nominative form. For the closed word classes, the base form does not differ from the full word form.

Morphemes are the smallest linguistic units carrying a semantic interpretation. A base morpheme gives a word its meaning. Affixes are morphemes attached to a base morpheme. Usually they cannot stand alone. The affix preceding a base morpheme is called a prefix, and the one coming after the base morpheme is called a suffix. Affixes can be inflectional or derivational. For example, the word "unbreakable" has three morphemes: "break" is the base morpheme, "un" is a derivational prefix and "able" is a derivational suffix. In the word "goes", "go" is the base morpheme and "es" is the inflectional suffix. The stem is a base morpheme which remains when all inflectional affixes have been removed. The root is a base morpheme which remains after removal of all affixes, both inflectional and derivational.

Part-of-speech (POS) tags, also called grammatical tags, classify words into categories (classes) depending on the context relationship with adjacent and related words in a phrase or sentence. Most POS tag sets make use of the same basic categories, such as nouns, verbs, adjectives, adverbs, pronouns, determiners, prepositions, conjunctions and numerals. However, the detailed tag sets can differ both in how finely they divide words into categories and in how they define the categories. For example, "is" might be tagged as a verb in the basic tag set, as a present tense of the verb in one detailed tag set, and as a third person singular present tense form of the verb in another detailed tag set. This variation in tag sets is unavoidable, since POS tags are used in different ways for different tasks. In other words, there is no "right way" to define tags, only more or less useful ways, depending on the tasks and goals.

3.2 Analysis and annotation

A prerequisite for the methods described in this work is the availability of morpho-syntactic analysers which can provide the translation corpora with appropriate morpho-syntactic information. The construction of such analysers is a demanding task in its own right. Usually, they are based either on pure linguistic concepts (e.g. constraint grammar parsers for English [Voutilainen 95] and German [Haapalainen & Majorin 95] or on a statistical approach (e.g. maximum entropy tagging [Ratnaparkhi 96], *n*-gram-based tagging [Brants 00], decision tree based analyser[1]). Each approach has advantages and disadvantages. The main advantage of the linguistic approach is that the analyser could be used on any text in the given language regardless of the domain. Another advantage is that no training material is needed. On the other hand, the statistical approach does not have problems with annotation of "ungrammatical" sentences which often appear in dialogues or outputs generated by speech recognition systems. In addition, they can be used for any language provided that a suitable annotated training corpus for the desired language can be found. Among the statistical approaches, the maximum entropy framework [Ratnaparkhi 96] as well as the *n*-gram models combined with a good smoothing technique and handling of unknown words [Brants 00] have a very strong position.

For the Spanish, German and English languages there are high quality analysers available. For the experiments in this thesis, the morpho-syntactic annotation of the English corpora is performed using the constraint grammar parser ENGCG [Voutilainen 95]. The German corpora are also annotated using the constraint grammar parser GERCG [Haapalainen & Majorin 95] and the Spanish texts are annotated using the FreeLing analyser [Carreras & Chao$^+$ 04]. In this way, all texts are provided with POS tags and base forms. For applications where only POS tags are needed (e.g. POS-based word reorderings, syntactic-oriented evaluation measures), the statistical *n*-gram-based TNT-tagger [Brants 00] can be also used for the English and German corpora. For the Serbian language, so far there are no morpho-syntactic analysers, and annotation of this corpus is done half manually and half automatically. All base forms have been introduced manually and the POS tags have been provided by an iterative procedure including manual annotation and the use of a statistical maximum-entropy based POS tagger [Bender 02] similar to the one described in [Ratnaparkhi 96]. All these tools have high accuracy of about 97% (i.e. low error rate of about 3%).

Certain types of morpho-syntactic information can be also extracted from the text itself without the help of any linguistic tools, such as the identification of compound words or word stems and suffixes. These corpus-based methods are based on the frequencies of certain words and their components in the corpus, like for example the splitting of German compounds presented in [Koehn & Knight 03]. The main advantage of such methods is that no external morpho-syntactic analyser for the desired language is needed. The main disadvantage is that they may provide some "non-linguistic" annotations as well as omit some linguistic ones. For example, the German word "Treibhauseffekt" will be identified as a compound only if both components "Treibhaus" and "Effekt" are seen in the corpus as single words. However, these phenomena are not crucial when the main goal is improving the quality of SMT.

[1] http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

4 POS-based word reorderings

As already mentioned in previous chapters, word reordering is an important issue in SMT. Although statistical word alignments work rather well at capturing differences in word order and a number of strategies for non-monotonic search have been developed, differences in word order between the source and the target language are still one of the main causes of translation errors.

In this work, we investigate possibilities for improving the translation quality by rule-based reordering of the source sentence using only POS tags. The source languages in our experiments are Spanish, German and English. Reorderings are applied in the source language, then training and search are performed using the transformed source language data. Modifications of the training and search procedure are not necessary. Figure 4.1 represents a general training and translation process with reordering transformations.

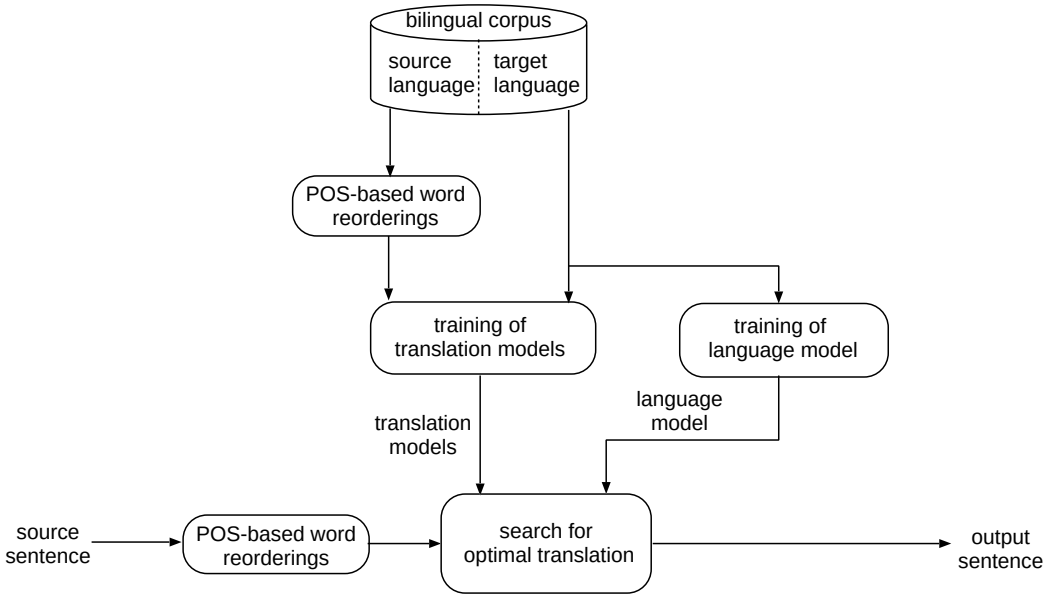

Figure 4.1: Training and translation with POS-based word reorderings as a preprocessing step.

We propose two types of reorderings depending on the language pair:

- local reorderings (convenient for translation from and into Spanish), and
- long-range reorderings (convenient for translation from and into German).

Local reorderings are applied on adjectives and long-range reorderings on verbs. All reorderings are done automatically.

In addition to the fixed rule-based reorderings of the test corpus, a translation of word graphs is also investigated, where each source text sentence is replaced by a word graph. This word graph contains all possible paths obtained by combining the original text with the reordered text. When the word graph is translated, transformations of the source part of the training corpus are possible but not mandatory. In this work both variants are investigated. The overall translation process for word graphs is illustrated in Figure 4.2.

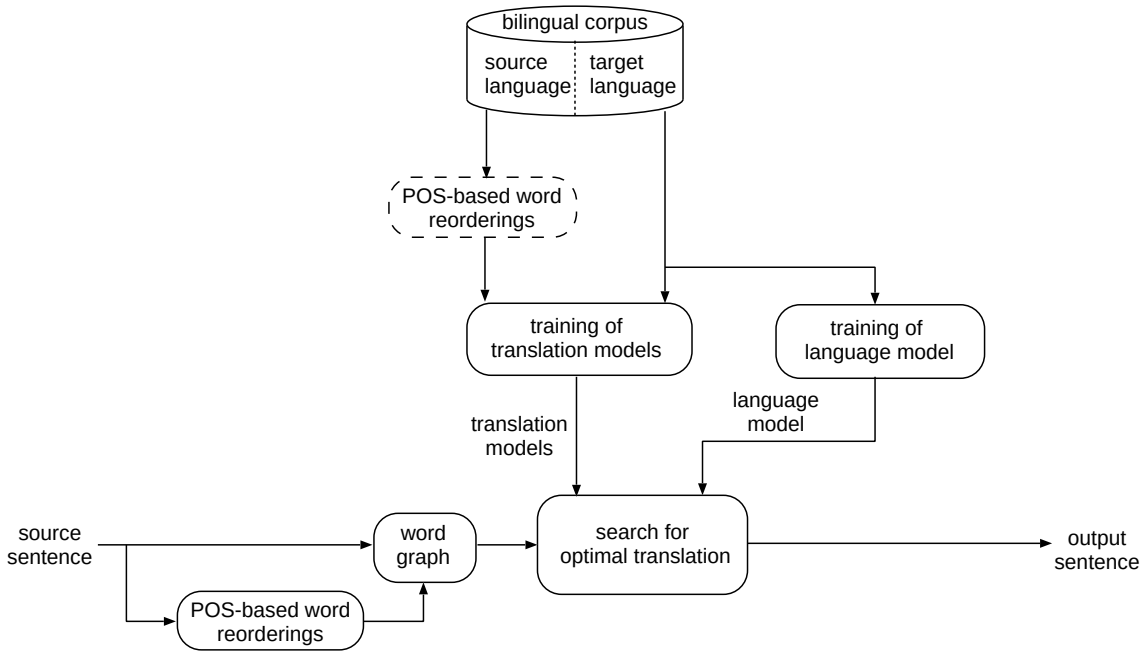

Figure 4.2: Translation of a word graph created from the original text in combination with the reordered text; reordering of the training corpus is possible but not mandatory.

The main motivation for investigating translation with word graphs is the fact that some applied reorderings might not be really appropriate due to POS tagging errors or to specific phenomena which are hard to cover with rules. An example is the Spanish–English language pair, where in English, adjectives always precede the corresponding nouns whereas in Spanish both variants are possible. Therefore some reorderings in English might not be optimal.

It should be noted that the word graphs described in this work, contrary to the majority of publications dealing with similar problems, do not contain any probabilities. The reason is that in this work the only source of information for reordering rules are POS tags in the source language, and the rules are based on the human knowledge about syntactic differences between involved languages. In other words, monolingual linguistic knowledge is the only source of information. On the other hand, in the majority of publications dealing with reorderings and word graphs, the rules are extracted bilingually, i.e. from the statistical word alignments in combination with POS tags so that frequencies of crossings of particular POS sequences in the alignment are used for estimating probabilities. For the methods described in this work, such an approach is of course not possible. Therefore we tried to extract probabilities from POS sequences in the target language in the following way: for example, for translation from Spanish into English we count the number of occurrences of the sequence “noun adjective” and the sequence “adjec-

tive noun" in the English training corpus, and divide each of these frequencies with their sum. However, preliminary experiments have not yielded any improvements in comparison with the word graph without probabilities, so we decided to let this aspect to be explored thoroughly in the future work.

4.1 Local reorderings

The local reorderings investigated in this work are applied on the Spanish–English language pair and handle differences between the positions of nouns and adjectives in the two languages. Adjectives in the Spanish language, as in most Romanic languages, are usually placed after the corresponding noun, whereas for English it is the other way round. Therefore, for this language pair local reorderings of nouns and adjective groups in the source language are applied. The following sequences of words are considered to be an adjective group: a single adjective, two or more consecutive adjectives ("difficult political" situation), a sequence of adjectives and coordinate conjunctions ("economic and political"), as well as an adjective along with its corresponding adverb ("more important"). If the source language is Spanish, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun.

Some examples of local reorderings of Spanish and English nouns and adjective groups can be seen in Table 4.1. In the first sentence the adjective group consists of two adjectives with a conjunction, and in the second one of an adjective and an adverb.

Table 4.1: Examples of local reorderings for Spanish and English nouns and adjective groups.

Spanish	motivos **económicos y políticos**	situación **claramente insatisfactoria**
	⇓	⇓
	económicos y políticos motivos	**claramente insatisfactoria** situación
English	**economic and political** reasons	**clearly unsatisfactory** situation
	⇓	⇓
	reasons **economic and political**	situation **clearly unsatisfactory**

An example of a Spanish word graph based on local reorderings is presented in Figure 4.3. In the case of hard preprocessing, only one (reordered) sentence is given to the decoder, namely "La claramente insatisfactoria situación de los humanos derechos...". When a word graph is translated, the decoder has four possibilities: the original sentence, the completely reordered sentence and two more alternative paths, "La claramente insatisfactoria situación de los derechos humanos..." and "La situación claramente insatisfactoria de los humanos derechos...".

4.2 Long-range reorderings

Long-range word reorderings in this work are tested on the verb groups for the German–English language pair. Verbs in the German language, unlike many other languages, can often

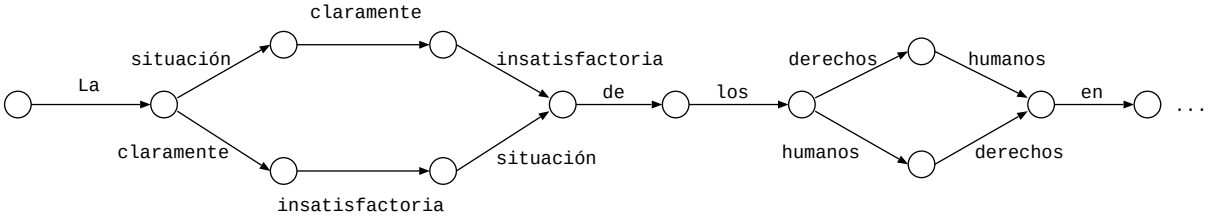

Figure 4.3: Example of a word graph created with local reorderings of nouns and adjective groups in the Spanish language.

be placed at the end of a clause. This is mostly the case with infinitives and past participles, but there are many cases when other verb forms also occur at the end of the clause. Therefore long-range reorderings of verb groups are appropriate for translation systems where the German language is involved.

In this work, we investigate both translation directions. For translation **from German into English** we investigate all types of long-range syntactic phenomena related with German verbs which lead to quite different word order in English. The following types of reorderings are applied:

infinitives: In the German language, infitinives appear at the end of the clause if an auxiliary or a modal verb is present. In English they appear immediately after the auxiliary/modal verb. Therefore each German infinitive (bold) is moved back to the closest auxiliary/modal verb (italic).

German	reordered German	English
es *wird* ein Kapitel über Wissenschaft **geben**	es *wird* **geben** ein Kapitel über Wissenschaft	there will be a chapter on science
ich *kann* es Ihnen heute **sagen**	ich *kann* **sagen** es Ihnen heute	I can tell you today

infinitives+zu: This construction always appears at the end of the clause, often together with the conjunction “um”. The English equivalent consists of the particle “to” and an infinitive, sometimes of the expression “in order to” and an infinitive.

Reordering is performed as follows: each infinitive+zu group (bold) is moved back to the beginning of the clause or to the closest conjunction “um” (italic).

German	reordered German	English
noch weiter **zu gehen**	**zu gehen** noch weiter	to go further
um einige Beispiele **zu zeigen**	*um* **zu zeigen** einige Beispiele	to show some examples
um EU Standards **zu erreichen**	*um* **zu erreichen** EU Standards	in order to attain EU standards

finite verbs: Some German subordinate conjunctions send finite verbs to the end of the subordinate clause, such as “dass” (that), “weil” (because), “damit” (so that), etc. If the finite verb is auxiliary or modal, the main verb is an infinitive or a past participle and the order of the verbs is inverted. Reordering rules are applied in the following way: if a sentence

contains a subordinate conjunction which affects the position of finite verbs (italic), the finite verbs (bold) are moved back next to the conjunction. If other verbs are present at the end of the clause, they are moved back to immediately follow the finite verb.

German	reordered German	English
dass ich zu einem anderen Schluss **komme**	*dass* ich **komme** zu einem anderen Schluss	that I have reached a different conclusion
weil wir Massengüter weg von der Straße **bekommen müssen**	*weil* wir **müssen bekommen** Massengüter weg von der Straße	because we must get heavy freight off the roads

In the first sentence, the finite verb "komme" is the main verb and the conjunction "dass" sends it to the end of the clause. In the second sentence, the finite verb is the modal verb "müssen" whereas the main verb "bekommen" is in the infinitive form. The conjunction "weil" sends the finite verb to the end of the clause so that the two verbs appear in reverse order.

past participles: Similarly to the infinitives, German past participles could also appear at the end of a clause. Therefore each past participle (bold) is moved back to the closest auxiliary verb (italic).

German	reordered German	English
ich *habe* für den Bericht **gestimmt**	ich *habe* **gestimmt** für den Bericht	I have voted for the report

negative particles: The negative particle in German often occurs relatively far away from the finite verb towards the end of a clause, in contrast to English where this particle is always close to the finite verb. Therefore the negative particles (bold) are moved back to the closest finite verb (italic).

German	reordered German	English
das *ist* in einem modernen Europa **nicht** möglich	das *ist* **nicht** in einem modernen Europa möglich	this is not possible in a modern Europe

prefix: Some German verbs consist of a main root and a separable prefix which can appear at the end of the clause. In addition, this prefix sometimes can be separated and sometimes attached to the main root. We investigate two possibilities to handle this phenomenon: to move the prefix (bold) to immediately precede the root (italic), or to join the prefix with the root.

German	reordered prefix	appended prefix	English
ich *gehe* davon **aus**	ich **aus** *gehe* davon	ich **ausgehe** davon	I assume
die Abstimmung *findet* heute um 12:30 Uhr **statt**	die Abstimmung **statt** *findet* heute um 12:30 Uhr	die Abstimmung **stattfindet** heute um 12:30 Uhr	the vote will take place today at 12.30h

In our experiments we explore the impact of each type of reordering separately, as well as the effects of applying all reorderings together. Examples of sentences which are affected with different types of reorderings are shown in Table 4.2. In the first sentence, the first clause with reorderings has the conjunction "dass" which sends the finite auxiliary verb "werden" to

the end. In addition, an infinitive of the main verb is present, so the finite verb is moved to the beginning of the clause, i.e. after the subordinate conjunction “dass”, and the infinitive is moved to the position after the auxiliary verb. The second clause contains an infinitive with “zu” – “vorzubereiten”, which is moved to the beginning of the clause. The rules for finite verbs are again applied to the third clause; the finite verb “gehört” is moved to follow directly the conjunction “wozu”. In the second sentence, the infinitive “fassen” is reordered to follow its auxiliary verb “werden” in the first clause. In the second clause two types of reorderings are present; the finite auxiliary verb “habe” is moved back to the subordinate conjunction “wie”, and the past participle “getan” is reordered immediately after the finite verb. The third clause has three reorderings; the finite auxiliary verb “habe” is moved to follow the conjunction “als”, the negative particle “nicht” is reordered to follow the finite verb, and then the past participle “genutzt” is moved to the auxiliary verb.

Table 4.2: Examples of long-range reorderings of different types of German verbs.

German:	Die Kommission ist der Auffassung, *dass*^{fin} alle Unternehmen in der Lage **sein**^{inf} **werden**^{fin}, sich rechtzeitig auf die endgültige Umstellung auf den Euro **vorzubereiten**^{infzu}, *wozu*^{fin} auch die Anpassung ihrer Computersysteme und der Software **gehört**^{fin}.
reordered:	Die Kommission ist der Auffassung, dass **werden sein** alle Unternehmen in der Lage, **vorzubereiten** sich rechtzeitig auf die endgültige Umstellung auf den Euro, wozu **gehört** auch die Anpassung ihrer Computersysteme und der Software.
English:	The Commission believes that all undertakings will be able to prepare themselves in time for the final changeover to the euro, including the adaptation of their computer systems and software.
German:	Ich *werde*^{inf} mich kurz **fassen**^{inf}, *wie*^{fin} ich dies schon vorhin **getan**^{part} **habe**^{fin}, *als*^{fin} ich bei meinem ersten Redebeitrag drei Minuten meiner Redezeit **nicht**^{neg} **genutzt**^{part} **habe**^{fin}.
reordered:	Ich werde **fassen** mich kurz , wie ich **habe getan** dies schon vorhin , als ich **habe genutzt nicht** bei meinem ersten Redebeitrag drei Minuten meiner Redezeit .
English:	I shall be brief , as indeed I was earlier , since my first speech was three minutes under the allotted speaking time .

As in the case of the Spanish–English pair, we also investigate the translation of word graphs. An example of a German word graph based on long-range reorderings of past participles is presented in Figure 4.4. As in the case of the local reorderings, the decoder can choose among various paths. For the graph in Figure 4.4 there are two possibilities: original sentence and reordered sentence. For the examples in Table 4.2 there will be eight possible paths – original sentence, reordered sentence and six other possibilities with one or two reordered clauses.

In addition to these graphs, for German→English translation we also explore word graphs with local reorderings within long-range reorderings. The main reason for this is the fact that the long-range reorderings are not always perfect in the local context. An example can be seen in the second sentence in Table 4.2: the reordering of the finite verb, the past participle and the

negative particle “habe genutzt nicht” in the local context is questionable; maybe the sequence “habe nicht genutzt” would be better for translation into English? Besides, the German language in general has more free word order than English allowing, for example, finite verbs to precede the subject pronouns, object pronouns to precede finite verbs and subject pronouns, etc. These also might lead to not completely appropriate long-range reorderings in the local context. An illustration is presented in Table 4.3: the reordering which perfectly matches the English word order would be “deswegen ich möchte vorschlagen Ihnen”.

Table 4.3: Example of differences between German and English word order in the local context: finite verbs can preceed the subject pronouns.

German:	deswegen möchte^{finV} ich^{subjPron} Ihnen^{objPron} vorschlagen^{mainV}
reordered:	deswegen möchte vorschlagen ich Ihnen
English:	that is why I should like to propose to you

Therefore, we investigate introducing addition local reorderings into the word graphs generated by long-range reorderings. The local reorderings are added for each sequence of verbs containing negative particles as well as for each sequence of consecutive pronouns and verbs. Word graphs for previously described examples are presented in Figure 4.5. As can be seen, some additional paths are added giving more possibilities to the decoder.

For translation **from English to German** we investigate two types of reorderings: infinitives and past participles. All infinitives and past participles are moved to the end of a clause where punctuation marks, subordinate conjunctions and finite verbs are considered as the beginning of the next clause.

infinitives: All infinitives are moved to the end of the clause. The infinitives following an auxiliary or modal verb are separated from this verb and moved. The infinitives with the preceding infinitive particle “to” are moved to the clause end together with the particle.

English:	We **have to offer** them our hand in a very symbolic way.
reordered:	We *have* them our hand in a very symbolic way **to offer**.
German:	Wir müssen dem Volk ganz symbolisch die Hand reichen.
English:	We *must* **do** everything possible **to douse** the fire as quickly as possible.
reordered:	We *must* everything possible **do** the fire as quickly as possible **to douse**.
German:	Es muss alles getan werden, um den Brand so schnell wie möglich einzudämmen.

past participles: All past participles are separated from its auxiliary verb and moved to the end of the clause.

English:	I *have* already **answered** this question.
reordered:	I *have* already this question **answered**.
German:	Ich habe diese Frage bereits beantwortet.

The translation of word graphs is explored for this translation direction too, and the word graphs are created as combination of original and reordered sentences in the same way as for the Spanish–English pair and for German–English translation without additional local context.

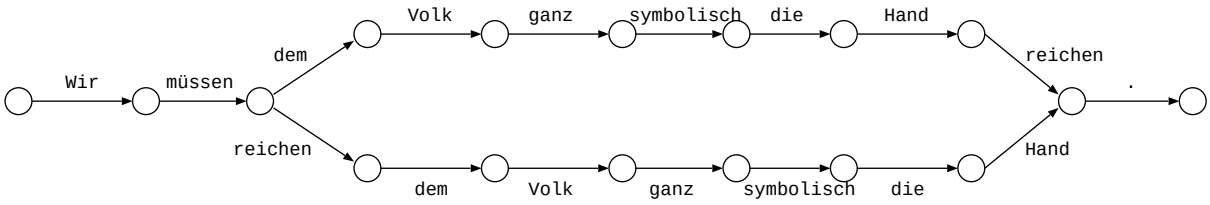

Figure 4.4: Example of a word graph created with long-range reorderings of verb infinitive in the German language.

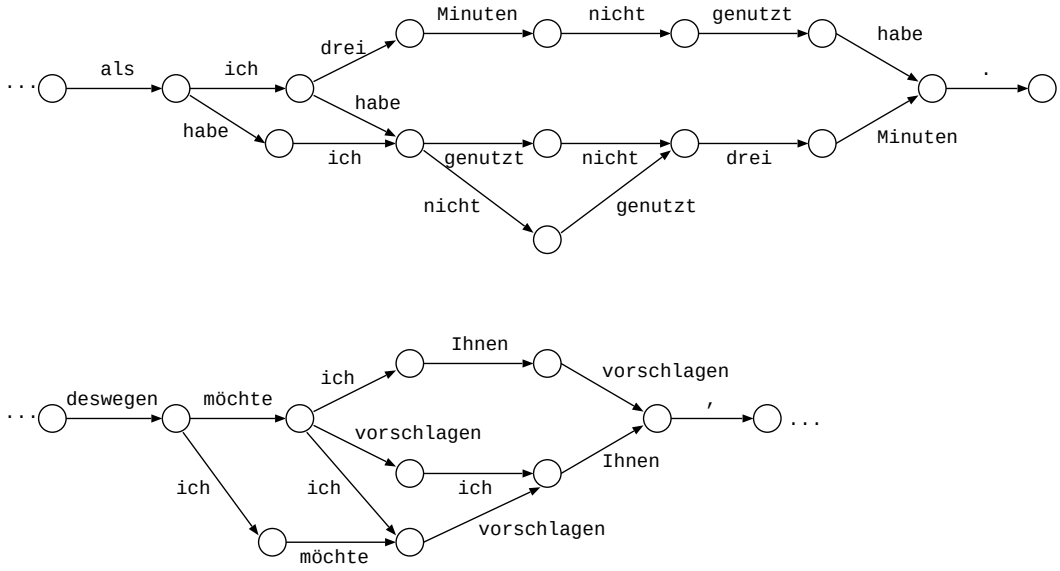

Figure 4.5: Examples of word graphs with local reorderings within the long-range reorderings in the German language.

4.3 Experimental results

Experimental settings

Baseline translation system: The baseline translation system is the phrase-based system similar to systems described in [Zens & Bender$^+$ 05, Matusov & Zens$^+$ 06]. The key elements of this translation approach are bilingual phrases, i.e. pairs of source and target language phrases, where a phrase is a contiguous sequence of words. These bilingual phrases are extracted from a word-aligned bilingual training corpus. GIZA++ is used to train this word alignment.[1] To obtain a more symmetric word alignment, the training is performed in both translation directions and the resulting Viterbi alignments are unified [Och & Ney 03].

The posterior probability $Pr(e_1^I|f_1^J)$ is modelled directly using a weighted log-linear combination of a trigram language model and various translation models: phrase translation models in source-to-target and target-to-source directions and word lexicon models similar to the IBM-1 model, also in both directions. Additionally, a word penalty and a phrase penalty are used. With the exception of the language model, all models can be considered as within-phrase models, because they depend only on a single phrase pair and not on the context outside the phrase. The language model is a trigram model [Stolcke 02] with modified Kneser-Ney discounting and interpolation [Kneser & Ney 95]. The seven submodels are combined by weighted log-linear interpolation. The scaling factors $\lambda_1, \ldots, \lambda_7$ are optimised with respect to the maximum BLEU score [Och 03] on a development set. The search procedure is monotone, i.e. without reordering constraints.

Corpora: The proposed local and long-range word reorderings are tested on the transcriptions of the European Parliament Plenary Sessions. The statistics can be found in Appendix A, Sections A.1 and A.2. The long-range reorderings for German as a source language are tested additionally on the VERBMOBIL corpus since the first experiments in this direction [Nießen & Ney 01a] were performed on this corpus.

Evaluation metrics: The performance measures used for the assessment of translation quality are the BLEU score, TER (translation edit rate), WER (word error rate), PER (position-independent word error rate) and CDER (edit distance which allows reordering of blocks). In addition, POSBLEU– syntactic BLEU score, i.e. the BLEU score calculated on POS tags instead of words, is also presented. This metric has shown high correlation with human judgments in some recent experiments. All evaluation metrics are summarised in Appendix B. More details about the standard measures can be seen in Section B.1, whereas the description and results of experiments dealing with correlations between the POSBLEU and the human scores adequacy and fluency are presented in Section B.2.

Local reorderings

Local reorderings are tested on the Spanish–English EPPS corpus developed in the framework of the TC-STAR project. The details about this corpus are presented in Section A.1. In this

[1] The GIZA++ toolkit for word alignment can be downloaded from
`http://www.fjoch.com/GIZA++.html`

section, the results for the test corpus used in the second TC-STAR evaluation are presented. In addition, results obtained on the test corpus from the first evaluation as well as on the additional Spanish Parliament test corpus can be found in Appendix C.

Reorderings are applied as a preprocessing step on the source side of the training corpus as well as on the test corpus. Table 4.4 presents the percentage of sentences in the training and test corpus which are actually reordered. It can be seen that more than 60% of sentences are affected by reordering transformations for both source languages.

Table 4.4: Percentage of reordered sentences in the training and the test part of the TC-STAR Spanish–English corpus: local reorderings of adjectives.

	train	test
Spanish	60.8%	62.8%
English	64.2%	71.9%

As mentioned previously, the baseline system used the monotone search procedure. Preliminary experiments during the second TC-STAR evaluation showed that the POS-based reorderings yielded better results than the non-monotonone search using local reordering constraints with the window size 3. The results on the development set for Spanish-to-English translation are shown in Table 4.5. Applying local reordering constraints on the POS-based reordered corpus lead to some further improvements, therefore this configuration was used for the final submissions in the second as well as in the third TC-STAR evaluation. The results on the test corpus for the five best-ranked systems of the second TC-STAR evaluation can be seen in Table 4.6.

Table 4.5: Translation results [%] for the Spanish→English development corpus: monotone and non-monotone search with and without the POS based local reorderings of adjectives.

Spanish→English	BLEU	TER	WER	PER	CDER
monotone search	50.5	37.2	40.2	28.5	35.9
+reorder adjectives	51.8	35.8	38.7	27.2	34.9
non-monotone search	50.8	36.6	39.6	27.7	35.4
+reorder adjectives	52.0	35.7	38.6	27.1	34.7

Table 4.6: Translation results [%] for the five best systems of the second TC-STAR evaluation for the Spanish→English test corpus.

system	BLEU	WER	PER
IBM	54.1	36.2	26.4
RWTH	53.1	37.1	26.9
UW	52.6	37.6	27.2
IRST	52.4	37.4	27.2
UPC	52.3	37.0	27.2

Spanish→English: Table 4.7 presents the results for the translation from Spanish to English. It can be seen that the local reorderings have increased the BLEU score and reduced all error rates except PER. This could be expected since the PER does not take the word order into account. The changes in evaluation metrics are not statistically significant, but they are consistent; the results on more test data are presented in Appendix C.2 and the same tendencies can be observed for all texts.

More details can be seen in Table 4.8. The test corpus was divided into two parts: one containing sentences that have been actually reordered (which is about 60% according to Table 4.4) and the other containing sentences which do not change. These two sets were evaluated separately for each translation system: baseline and with local reorderings of adjectives. Results show that the reorderings improve translation quality of the reordered set where the rest remains basically the same.

Two translation examples are presented in Table 4.9. It can be seen that the system trained on the reordered corpus is more capable of producing the correct output; in the first sentence only the word order in the baseline output is incorrect and the local reorderings solved this problem. In the second sentence not only the word order but also the semantic meaning were not conveyed appropriately using the baseline system, whereas the output of the system with local reorderings is completely correct.

Table 4.7: Translation results [%] for Spanish→English: local reorderings of adjectives.

Spanish→English	BLEU	TER	WER	PER	CDER
baseline: monotone search	52.0	34.6	36.9	26.3	33.7
reorder adjectives	52.5	34.4	36.7	26.3	33.4

Table 4.8: Separated translation results [%] for Spanish→English: reordered sentences and the rest.

Spanish→English		BLEU	TER	WER	PER	CDER
reordered	baseline: monotone search	52.4	34.6	37.0	25.9	33.7
	reorder adjectives	53.4	34.2	36.6	25.9	33.1
not reordered	baseline: monotone-search	50.4	34.6	36.4	27.8	33.6
	reorder adjectives	50.2	34.6	36.3	27.8	33.7

English→Spanish: The results for this translation direction can be seen in Table 4.10. The effects are very similar as for Spanish→English translation. The improvements are again not statistically significant, but they are consistent (see Appendix C.2). The results of the separated evaluation shown in Table 4.11 are also similar as those for Spanish→English, except that the "unreordered" set in this case is slightly improved as well.

The translation examples in Table 4.12 show the advantages of the new system. In the first sentence, only the word order does not match the reference translation in the baseline output, whereas in the second sentence the wrong lexical choice and unclear semantic meaning are also present. It should be noted that the second sentence translated with the new system is still not

Table 4.9: Translation examples for Spanish→English translation with and without local reorderings of adjectives.

original Spanish sentence:	se trata de un *programa ambicioso y realista*
reordered Spanish sentence:	se trata de un *ambicioso y realista programa*
generated English sentence:	
without reordering:	it is of an **ambitious programme and realistic**
with reordering:	this is an **ambitious and realistic programme**
reference English sentence:	this is an ambitious and realistic programme
original Spanish sentence:	un intercambio de *experiencias prácticas*
reordered Spanish sentence:	un intercambio de *prácticas experiencias*
generated English sentence:	
without reordering:	an exchange of **experiences and practices**
with reordering:	an exchange of **practical experience**
reference English sentence:	an exchange of practical experiences

syntactically and morphologically perfect; the second “que” is inserted and the verb inflection “ha” is incorrect. However, the meaning is conveyed correctly.

The syntactic evaluation measure POSBLEU for the TC-STAR corpus is presented in Table 4.13. The results indicate that the syntactic structure is improved by local reorderings for both target languages. Separated POSBLEU scores are shown in Table 4.14. As the standard evaluation measures, the results of the reordered sentences are improved substantially whereas for the rest of sentences there are no significant changes for the English output and small improvements can be observed for the Spanish output.

Table 4.10: Translation results [%] for English→Spanish: local reordering of adjectives.

English→Spanish	BLEU	TER	WER	PER	CDER
baseline: monotone search	48.2	38.5	40.4	31.1	37.6
reorder adjectives	49.0	38.1	39.8	30.8	37.0

Table 4.11: Separated translation results [%] for English→Spanish: reordered sentences and the rest.

English→Spanish		BLEU	TER	WER	PER	CDER
reordered	baseline: monotone search	48.4	38.6	40.6	30.7	37.8
	reorder adjectives	49.3	38.1	40.0	30.4	37.3
not reordered	baseline: monotone search	47.2	38.3	39.3	33.1	36.6
	reorder adjectives	47.7	37.9	38.9	32.7	35.9

Table 4.12: Translation examples for English→Spanish translation with and without local reorderings of adjectives.

original English sentence:	a *legally binding ban*
reordered English sentence	a *ban legally binding*
generated Spanish sentence:	
without reordering:	un **jurídicamente vinculante prohibición**
with reordering:	una **prohibición jurídicamente vinculante**
refrence Spanish sentence:	una prohibición legalmente vinculante
original English sentence:	we have acknowledged that *positive change* has occurred
reordered English sentence:	we have acknowledged that *change positive* has occurred
generated Spanish sentence:	
without reordering:	hemos reconocido que **positivo ha cambiado**
with reordering:	hemos reconocido que **cambios positivos** que ha ocurrido
reference Spanish sentence:	hemos reconocido que se han registrado cambios positivos

Table 4.13: POSBLEU scores [%] for both translation directions on the TC-STAR corpus: local reorderings of adjectives.

	Spanish→English	English→Spanish
baseline: monotone search	68.2	57.5
reorder adjectives	68.9	58.3

Table 4.14: Separated POSBLEU scores [%] for both translation directions on the TC-STAR corpus: reordered sentences and the rest.

		Spanish→English	English→Spanish
reordered	baseline: monotone search	68.5	57.9
	reorder adjectives	69.6	58.9
not reordered	baseline: monotone search	66.7	55.8
	reorder adjectives	66.5	56.5

Translation of word graphs

The translation of the word graphs described in Section 4.1 is performed in both translation directions. The word graphs both for the Spanish and English test corpora are created by combining the original and locally reordered sentences, thus containing all possible paths. For training, two possibilities are explored: training with the original corpus (referred to in the tables as "baseline") and training with the reordered source language corpus (referred to as "reorder adjectives").

The results for Spanish→English translation are reported in Table 4.15. When the system is trained on the original corpus, the use of a word graph improves all error rates except PER. However, if the reordered training corpus is used, the improvements for some error rates are larger. For this training, using a word graph instead of reordered test corpus does not yield any improvements. A probable reason is the characteristic of the English language mentioned at the beginning of this chapter – the adjective should always precede the noun. Therefore after the fixed reordering of the Spanish corpus there are not many ambiguities left which could be resolved by word graphs.

Table 4.15: Translation of word graphs: results [%] for Spanish→English translation of the TC-STAR corpus.

Spanish→English	BLEU	TER	WER	PER	CDER
baseline: monotone search	52.0	34.6	36.9	26.3	33.7
+graph	52.3	**34.4**	**36.7**	26.3	33.6
reorder adjectives	**52.5**	**34.4**	**36.7**	26.3	**33.4**
+graph	52.3	34.5	36.8	26.3	33.5

Table 4.16 shows the results for the translation from English into Spanish. As in the case of the other translation direction, the use of a graph with the baseline system results in improvements for all evaluation metrics. However, in contrast to translation from Spanish, training with the reordered corpus yields the same improvements as with the original corpus. Furthermore, translating word graph with the new system (trained on reordered corpus) results in some additional improvements.

Table 4.16: Translation of word graphs: results [%] for English→Spanish translation of the TC-STAR corpus.

English→Spanish	BLEU	TER	WER	PER	CDER
baseline: monotone search	48.2	38.5	40.4	31.1	37.6
+graph	49.0	38.0	39.8	30.8	37.0
reorder adjectives	49.0	38.1	39.8	30.8	37.0
+graph	**49.3**	**37.8**	**39.7**	**30.7**	**36.8**

A probable reason for the different tendencies is the characteristic of the Spanish language described at the beginning of the chapter. Although in Spanish the adjective groups *usually* follow the corresponding noun, this is not *always* the case. However, there are no straightforward

rules to determine which variant is preferred in which case. Therefore some fixed reorderings in English are not really helpful, whereas the word graph allows the decoder to choose the optimal word sequence.

Long-range reorderings

long-range reorderings are tested on the German–English EUROPARL corpus developed for the second and third shared task of the Statistical Machine Translation Workshop and on the VERBMOBIL corpus. More details about the corpora along with the corpus statistics can be found in Section A.2. Like the local reorderings described in the previous section, these reorderings are also applied as a preprocessing step both for the training and for the translation.

For translation from German into English, the reorderings described in Section 4.2 are tested both separately and combined. The percentage of sentences affected by each reordering is presented in Table 4.17 for both corpora. In the EUROPARL corpus, about half of the total number of sentences contain finite verb reordering. Infinitives, infinitives with "zu" and past participles are applied in about 20% of sentences, negative particle reordering occurs in about 10% sentences, and only in 7% of sentences the verb prefix is reordered. For the VERBMOBIL corpus, the distribution of reorderings is different: in 30% of sentences infinitives are reordered, past participles in about 15%, followed by prefixes, negative particles and finite verbs with about 6% of affected sentences, whereas reordering of infinitives with "zu" occurs in less than 1% of sentences. When all reorderings are applied, between 70 and 80% of sentences are affected in both test corpora.

For translation from English into German, only the combined reorderings of infinitives and past participles are tested. About 60% of the sentences are affected by these reorderings. In future, more reorderings for this translation direction should be investigated and tested both separately and combined.

Table 4.17: Percentage of reordered sentences: German sentences in the EUROPARL and VERBMOBIL corpus, English sentences in the EUROPARL corpus.

German	EUROPARL		VERBMOBIL	
	train	test	train	test
infinitive	25.6%	29.5%	12.6%	32.3%
infinitive+zu	15.6%	20.6%	1.7%	0.8%
finite	44.7%	52.8%	7.7%	6.4%
past participle	33.4%	21.2%	3.2%	15.5%
negative particle	10.5%	11.0%	3.1%	6.8%
prefix	6.9%	7.6%	4.8%	8.8%
all	92.6%	79.8%	49.9%	74.0%

English, EUROPARL	train	test
infinitive+past participle	54.3%	61.6%

As already mentioned at the beginning of the section, the monotone search procedure is used for the baseline system. Preliminary experiments with the IBM reordering constraints showed only minor improvements (0.1% absolute for the BLEU score) so that only monotone search and POS-based reorderings are presented in this work.

German→English: As mentioned above, for this translation direction each reordering is tested separately as well as in combination with other reorderings. The following combinations are tested:

- infinitives, negative particles, prefixes and finite verbs as proposed in [Collins & Koehn$^+$ 05] (inf+neg+pref+fin),
- the most often reorderings in the EUROPARL corpus, i.e. infinitives with and without "zu", finite verbs and participles (infzu+inf+fin+part),
- the combination above together with negative particles (+neg),
- the combination above with
  - reordering of prefixes (+pref),
  - appending prefixes (+append pref).

Two ways of prefix treatment are explored: reordering, i.e. moving the prefix to precede the root as proposed in [Collins & Koehn$^+$ 05], and appending, i.e. joining the prefix with the root as proposed in [Nießen & Ney 01a].[2]

Table 4.18 presents the results of the different types of long-range verb reorderings for the EUROPARL corpus. The BLEU score of the baseline system (24.4%) is comparable with the BLEU scores obtained on the same test corpus by the two best-ranked systems (24.7% and 24.3%) at the second shared task on statistical machine translation [Koehn & Monz 05]. The largest improvements are obtained by combining the most frequent reorderings together with the negative particles. This method includes the novel reorderings proposed in this work, i.e. treatment of infinitives with "zu" and past participles. As for the separated reorderings, the best improvements are achieved with infinitives with "zu" and finite verbs.

Contrary to the results presented in [Nießen & Ney 01a] on the VERBMOBIL corpus, treatment of verb prefixes did not yield any improvements for the EUROPARL corpus, neither as the only transformation nor in combination with other reorderings. That was one motivation to perform the same experiments on the VERBMOBIL corpus, and these results are presented in Table 4.19.

It can be seen that both treatments of the verb prefix improve translation quality, and that appending verb prefixes to the root result in better results than reordering them. However, the largest separate improvements are achieved with reorderings of finite verbs, and the best improvements are obtained by the same method as for the EUROPARL corpus, namely by reordering infinitives with and without "zu", past participles, finite verbs and negative particles. Adding verb prefix treatment to this configuration slightly lowers the results.

Separated results for reordered sentences and for the rest are shown in Table 4.20 for the EUROPARL corpus and in Table 4.21 for the VERBMOBIL corpus. It can be noted that for the

[2] It should be noted that the numbers are not the same as those reported in [Nießen & Ney 01a] and [Collins & Koehn$^+$ 05] due to differences between alignment training, translation system as well as between evaluation tools.

Table 4.18: Translation results [%] for German→English translation of the EUROPARL corpus for different types of long-range verb reorderings.

German→English, EUROPARL	BLEU	TER	WER	PER	CDER
baseline: monotone search	24.4	61.3	66.9	45.9	55.6
reorder infinitive	24.5	61.1	66.8	45.6	55.5
infinitive+zu	*24.9*	*60.9*	66.5	*45.7*	*55.1*
finite	24.7	*60.9*	*66.3*	45.8	*55.1*
past participle	24.8	61.2	66.8	45.9	55.3
negative particle	24.4	61.3	66.9	*45.7*	55.6
prefix	24.4	61.3	66.9	45.8	55.6
append prefix	24.5	61.3	66.9	45.8	55.5
reorder inf+neg+pref+fin	24.9	60.5	65.9	45.6	55.0
reorder infzu+inf+fin+part	25.4	**60.0**	**65.2**	**45.3**	**54.4**
+neg	**25.6**	60.2	65.4	45.7	**54.4**
+pref	25.3	60.2	65.4	45.7	54.4
+append pref	25.3	60.2	65.4	45.7	54.4

Table 4.19: Translation results [%] for German→English translation of the VERBMOBIL corpus for different types of long-range verb reorderings.

German→English, VERBMOBIL	BLEU	TER	WER	PER	CDER
baseline: monotone search	38.4	37.5	43.1	27.5	38.3
reorder infinitive	39.5	37.1	42.2	27.5	37.8
infinitive+zu	39.2	36.8	42.9	*26.6*	37.8
finite	*39.7*	*36.5*	*42.4*	*26.6*	38.0
past participle	39.2	37.3	43.0	*26.6*	37.9
negative particle	38.7	37.1	43.1	27.0	38.0
prefix	39.0	37.1	42.7	27.2	*37.5*
append prefix	39.3	36.6	42.6	*26.6*	38.3
reorder inf+neg+pref+fin	39.3	37.0	42.1	27.8	38.0
reorder infzu+inf+part+fin	41.3	36.5	41.0	27.1	37.0
+neg	**41.8**	**35.4**	40.1	**26.4**	**36.5**
+pref	41.1	36.7	41.1	**26.4**	36.9
+append prefix	41.7	35.7	**40.0**	26.5	36.6

reordered set the error rates are significantly higher than for the other set. The combination of verb reorderings which showed the best results for both corpora is referred to as “reorder verb”. The results indicate that the long-range reorderings considerably improve translation quality both for reordered sentences and for the others. This means that the new system allows better learning of various phrases so that the translation quality has been improved both directly as well as indirectly.

Two translation examples from the EUROPARL corpus are presented in Table 4.22. In the first sentence translated without reorderings, the English verb is positioned at the end of the sentence

and conjunction “that” is inserted. When reorderings are applied, the translation output is completely correct. The second sentence translated by the baseline system is both syntactically and semantically incorrect, and with verb reorderings both problems are solved.

Table 4.20: Separated translation results [%] for German→English translation of the EUROPARL corpus: reordered sentences and the rest; “reorder verbs” denotes the configuration “infzu+inf+part+fin+neg”.

German→English, EUROPARL		BLEU	TER	WER	PER	CDER
reordered	baseline: monotone search	22.4	62.9	69.0	46.7	57.3
	reorder verbs	23.5	61.8	67.3	46.5	55.9
not reordered	baseline: monotone search	39.2	49.9	53.0	40.7	43.8
	reorder verbs	40.3	48.8	52.0	39.9	43.6

Table 4.21: Separated translation results [%] for German→English translation of the VERBMOBIL corpus: reordered sentences and the rest; “reorder verbs” denotes the configuration “infzu+inf+part+fin+neg”.

German→English, VERBMOBIL		BLEU	TER	WER	PER	CDER
reordered	baseline: monotone search	37.5	38.2	43.8	27.7	38.9
	reorder verbs	41.4	35.8	40.1	26.5	36.6
not reordered	baseline: monotone search	42.3	35.1	40.4	26.6	35.9
	reorder verbs	42.8	33.9	40.1	26.1	35.6

English→German: The translation results for English→German are presented in Table 4.23, where it can be seen that the long-range reorderings improve translation quality also for this translation direction. Reordering of infinitives and past participles is denoted as “reorder verbs”. The improvements are less significant than for German→English at least for three reasons: because translation into German is generally more difficult due to its rich morphology, because only two types of reorderings are applied and few more are possible, and because this kind of long-range reordering (i.e. moving verbs far from their corresponding auxiliaries/pronouns/etc) in a way makes learning of models more difficult since more unseen or rarely seen patterns are produced.

Results for the separated test corpus are presented in Table 4.24. As for the other translation direction, the error rates of the reordered set are significantly higher. The improvements from verb reorderings are notable both for the reordered set and for the rest, although for the reordered sentences they are significantly higher.

The translation examples in Table 4.25 show an improvement in the verb group translation. In the first sentence, without reorderings the English word order is present in the German output. In the second sentence without reordering, the main verb “sein” is missing, whereas the new system translates the whole verb group correctly.

Table 4.26 presents POSBLEU scores for the EUROPARL corpus. long-range verb reorderings improve the syntactic structure of both translation outputs, especially for translation from

Table 4.22: Examples of German→English translation of the EUROPARL corpus with and without long-range reorderings of verbs.

original German sentence:	Es ist an der Zeit, die Verträge *zu überarbeiten.*
reordered German sentence:	Es ist an der Zeit, *zu überarbeiten* die Verträge.
generated English sentence:	
without reordering:	It is time **that** the Treaties **to review**.
with reordering:	It is time **to revise** the Treaties.
reference English sentence:	It is time to review the Treaties.
original German sentence:	Zypern *wird* gleichsam als Brücke zu den Ländern dieser Region *fungieren* .
reordered German sentence:	Zypern *wird fungieren* gleichsam als Brücke zu den Ländern dieser Region .
generated English sentence:	
without reordering:	Cyprus **is almost** as a bridge to the countries in the region **act.**
with reordering:	Cyprus **will act** as a bridge to the countries of the region.
reference English sentence:	Cyprus will be a sort of bridge with the countries in the area.

Table 4.23: Translation results [%] for English→German translation of the EUROPARL corpus: long-range verb reorderings; "reorder verbs" denotes the reordering of infinitives and past participles.

English→German, EUROPARL	BLEU	TER	WER	PER	CDER
baseline: monotone search	18.2	68.6	72.6	54.2	61.3
reorder verbs	18.4	67.8	71.6	53.8	61.0

German into English. In Table 4.27 the separated POSBLEU scores can be seen. Improvements in the syntactic structure on the reordered set are substantial for both translation directions. The rest of the sentences in the English output are also improved, although much less. However, the rest of the sentences in the German output are slightly deteriorated by reorderings. One possible reason for this has already been mentioned in the discussion about the standard evaluation

Table 4.24: Separated translation results [%] for English→German translation of the EUROPARL corpus: reordered sentences and the rest.

English→German, EUROPARL		BLEU	TER	WER	PER	CDER
reordered	baseline: monotone search	15.8	70.6	75.0	55.0	63.1
	reorder verbs	16.1	69.6	73.8	54.7	62.8
not reordered	baseline: monotone search	23.4	63.8	66.9	52.2	56.9
	reorder verbs	23.6	63.4	66.5	51.8	56.9

Table 4.25: Examples of English→German translation of the EUROPARL corpus with and without long-range reorderings of verbs.

original English sentence:	I would urge you *to support* the relevant amendments.
reordered English sentence:	I would urge you the relevant amendments *to support*.
generated German sentence:	
without reordering:	Ich bitte Sie **zu unterstützen,** die entsprechenden Änderungsanträge .
with reordering:	Ich bitte Sie , die entsprechenden Änderungsanträge **zu unterstützen**.
reference German sentence:	Ich bitte um Unterstützung der entsprechenden Anträge .
original English sentence:	But here again we should *be* frank.
reordered English sentence:	But here again we should frank *be*.
generated German sentence:	
without reordering:	Auch hier sollten wir ehrlich.
with reordering:	Doch auch hier sollten wir offen **sein**.
reference German sentence:	Aber auch hier sollten wir offen sein.

metrics, namely that these long-range reorderings produce some difficulties for the translation models by introducing more (rarely seen) phrases.

Table 4.26: POSBLEU scores [%] for both translation directions on the EUROPARL corpus: long-range reorderings of verbs.

EUROPARL	German→English	English→German
baseline: monotone search	36.3	20.8
reorder verbs	38.2	21.2

Table 4.27: Separated POSBLEU scores [%] for both translation directions on the EUROPARL corpus: reordered sentences and the rest.

EUROPARL		German→English	English→German
reordered	baseline: monotone search	34.7	18.5
	reorder verbs	36.6	19.3
not reordered	baseline: monotone search	47.0	25.9
	reorder verbs	47.8	25.6

Translation of word graphs

For both translation directions, word graphs created from the original and the reordered sentences are investigated. As for the local reorderings described in Section 4.1, two training

options are explored: with the original corpus (baseline) and with the reordered source training corpus (reorder verb). For German→English translation, the additional paths introduced by local reorderings of different verb types and pronouns described in Section 4.2 are tested. The results for German→English translation are shown in Table 4.28.

For both corpora similar tendencies can be observed. The best results are obtained by training with the reordered German text and translation of word graphs based on long-range reorderings together with additional local reorderings. For the EUROPARL corpus, the translation of word graphs with the baseline system does not result in any improvements except of a small increase of the BLEU score. With the system trained on the reordered corpus, word graphs based on long-range reorderings perform better than simple reordered sentences, and additional improvements are achieved by introducing local reorderings in the graph. For the VERBMOBIL corpus, translation of a word graph with the baseline system already brings significant improvements. One possible reason for this is the different nature of the corpora; the VERBMOBIL corpus is small compared to the EUROPARL corpus, and sentences are shorter. Preprocessing of the training and test corpus achieve larger improvements which are further increased using word graphs, especially with graphs enhanced with local reorderings. These local reordering paths seem to have much more importance for the VERBMOBIL data, probably because of a number of interrogative sentences; as reported in [Nießen & Ney 01a], question inversion is an important issue for this corpus.

For the other translation direction, the results are reported in Table 4.29. Using word graphs already improves the translation quality slightly with the baseline system, but improvements are larger if the reordering of verbs is applied in training and test. Word graphs translated with the new system do not lead to any improvements. One possible reason is complexity of this kind of reorderings: after applying the fixed reorderings, there are still a number of possiblbe reorderings which cannot be covered with a simple word graph with two alternative paths.

Table 4.28: Translation of word graphs: results [%] for German→English translation of the EUROPARL and VERBMOBIL corpus.

German→English		BLEU	TER	WER	PER	CDER
EUROPARL	baseline: monotone search	24.4	61.3	66.9	45.9	55.6
	+graph	24.6	61.3	66.9	45.9	55.6
	reorder verbs	25.6	60.2	65.4	45.7	54.4
	+graph	25.8	60.0	65.3	45.7	54.3
	+local	**25.9**	**59.8**	**65.0**	**45.6**	**54.1**
VERBMOBIL	baseline: monotone search	38.4	37.5	43.1	27.5	38.3
	+graph	40.2	36.9	41.8	27.8	37.5
	reorder verbs	41.8	35.4	40.1	**26.4**	36.5
	+graph	41.9	35.2	40.4	26.7	36.2
	+local	**43.2**	**34.9**	**39.5**	26.6	**35.8**

Table 4.29: Translation of word graphs: results [%] for English→German translation of the EUROPARL corpus.

English→German	BLEU	TER	WER	PER	CDER
baseline: monotone search	18.2	68.6	72.6	54.2	61.3
+graph	**18.4**	68.2	72.1	54.0	61.0
reorder verbs	**18.4**	**67.8**	**71.6**	53.8	**61.0**
+graph	**18.4**	**67.8**	71.7	**53.7**	**61.0**

4.4 Conclusions

Two novel methods for harmonising the word order between source and target language using only POS tags of the source language are presented and systematically evaluated: local reorderings of adjectives and long-range reorderings of verbs. Experiments showed that both types of POS-based word reordering improve translation quality for different language pairs and translation directions. However, it should be kept in mind that the appropriate methods and the achieved improvements significantly depend on the language pair, the translation direction and the nature of the corpus.

Local reorderings are tested for translation between Spanish and English, and small but consistent improvements are observed for both translation directions on three test corpora. For the translation from Spanish into English, the BLEU score is increased from 52% to 52.5% and TER is reduced from 34.6% to 34.4%. For the other translation direction the BLEU score is increased from 48.2% to 49.0% and TER is reduced from 38.5% to 38.1%. These reorderings can be useful for any other language pair with similar discrepancies between word order, such as translation between any Romance language and (say) English or German. The method can be applied for any other word classes, such as verbs and adverbs, pronouns and verbs, etc.

long-range reorderings are examined for the translation between German and English. For translation from German into English, systematic experiments are performed on two different corpora in order to investigate each type of word reordering separately as well as their most promising combinations. The best combination yielded in increase of the BLEU score from 24.4% to 25.6% for the EUROPARL corpus and from 38.4% to 41.8% for the VERBMOBIL corpus. TER for the EUROPARL corpus is reduced from 61.3% to 60.0%, and for the VERBMOBIL corpus from 37.5% to 35.4%. Two novel verb reorderings are proposed which have never been dealt with in previous work, i.e. treating infinitives with “zu” and past participles, and are shown to be important for translation performance. Thus, all possible long-range discrepancies between German and English verbs are covered in this work. Improvements obtained for the other translation direction are smaller, partly because translation into German is difficult in general, and partly because only two possible reorderings are tested. The BLEU score is improved from 18.2% to 18.4% and TER from 68.6% to 67.8%. These experiments should be extended in future work by introducing more reorderings and systematic investigation of each one.

In addition to fixed reorderings of the test corpus, translation of word graphs allowing all possible paths produced with and without reordering is explored. Word graphs are translated with the baseline system without reorderings as well as with the new system trained on the

reordered source corpus. The following tendencies are observed:

- word graphs translated with the new system obtained the best results for German→English and English→Spanish translation;
- for German→English translation, additional paths generated by local reorderings of verbs and pronouns yielded best results for both corpora, especially for the VERBMOBIL corpus;
- fixed reorderings both in the training and in the test corpus produced the best results for Spanish→English and English→German translation.

The word graphs investigated in this work do not contain any probabilities. Standard probabilities presented in other publications are not appropriate for these methods since they are extracted from word alignments, and preliminary experiments using relative frequencies of particular POS tag sequences in the target language have not shown any improvements. For the methods described in this work, systematic investigations are necessary to determine the optimal way for calculating probabilities. They will be explored in the future.

5 Translation with scarce bilingual resources

The performance of a statistical machine translation system depends on the size of the available training corpus. Usually, the larger the corpus, the better is the performance of a corpus-based translation system. Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, the acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and for some language pairs is not even possible. In addition, small corpora have certain advantages: the possibility of the manual creation of the corpus, the possible manual corrections of an automatically collected corpus, low memory and time requirements for the training of a translation system, etc.

This work aims at achieving the best possible translation quality with the smallest possible amount of bilingual training data. Three language pairs are investigated using various types of training data and appropriate morpho-syntactic transformations. The general scheme for training and translation with scarce bilingual resources with the help of an additional morpho-syntactic knowledge is presented in Figure 5.1. The translation models are trained on the bilingual corpus. Applying adequate morpho-syntactic transformations on the source, target or both parts of this corpus can enable better learning of models from sparse data. The language model is trained on a monolingual corpus in the target language. This corpus can be the target part of the bilingual corpus, but can also be a totally independent text. Of course, the translation quality will depend on the nature of this text, so if possible, it should be related to the domain of the bilingual training material. For the translation process, the same morpho-syntactic transformations should be applied on the test corpus. Inverse transformations are necessary after the search if the target part is affected by morpho-syntactic modifications. For example, if we use splitting German compound words for translation from English into German, after the translation process we need to merge the split German components.

5.1 Bilingual corpora

5.1.1 Conventional dictionaries

The use of conventional dictionaries (one word and its translation(s) per entry) have been proposed in [Brown & Della Pietra$^+$ 93] and they are shown to be valuable resources for statistical machine translation systems. They can be used to augment and also to replace the training corpus. This thesis investigates both of those two aspects for two language pairs, Spanish–English and German–English. For each language pair, adequate morpho-syntactic information is used in order to enable better learning from the dictionaries.

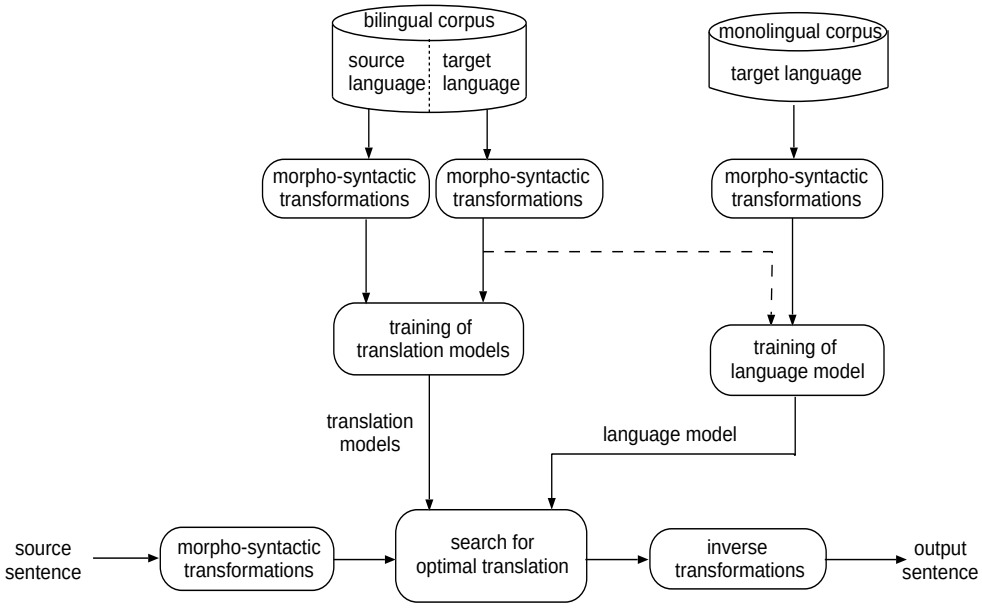

Figure 5.1: General scheme for training and translation with scarce bilingual resources and additional morpho-syntactic knowledge.

For the Spanish–English pair, a manually created conventional Spanish–English dictionary[1] not related to any particular task is used. The dictionary contains about fifty thousand entries, sixty thousand running words and thirty thousand distinct words for each language. This dictionary mainly contains short entries, i.e. single words and relatively short phrases and expressions. The average length is 1.2 words per entry. The majority of words are in their base form, but some inflected forms can be found too. More than 60% of the distinct words are singletons. The summarised statistics along with the test corpora are presented in Table A.1 in Appendix A.

The German–English dictionary used in this work is constructed by the concatenation of the dictionary created at Chemnitz university[2] and the dictionary used as additional bilingual knowledge source for the VERBMOBIL project described in [Nießen 02]. This dictionary consists mainly of short entries containing single words or small word groups, like the Spanish–English dictionary. On the other hand, as well as short entries, the Chemnitz dictionary contains a number of expressions in the form of complete (short) sentences thus covering a certain number of morphological and syntactic phenomena. The average length is about 2 words per entry. This dictionary has been manually created during more than ten years and new entries are still being added. The final dictionary consists of about three hundred thousand entries, four hundred thousand running German words and five hundred thousand running English words. The German vocabulary has more than one hundred thousand distinct words and the English about eighty thousand. About 60% of the distinct German words and 40% of the distinct English words are singletons. The complete statistics can be found in Table A.2.

[1] `http://www.quassa.com/lexiconData/english-spanish/eng-spa.dict.gz`
[2] `http://wftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/de-en.txt`

5.1.2 Phrasal lexica

Although conventional dictionaries are shown to be very useful, the main drawback is that they typically do not cover morphological and syntactic phenomena of languages. The entries normally contain only one or two words, usually base forms and not many inflections. The use of morphological expansions for overcoming morphological problems is investigated in [Nießen & Ney 04] for translation from German into English and in [Vogel & Monson 04] for translation from Chinese into English. Still, the dictionaries normally contain one word per entry and do not take into account phrases, idioms and similar complex expressions.

This work, in addition to dictionaries, exploits a phrasal lexicon (one phrase and its translation(s) per entry) as a bilingual knowledge source for SMT. A phrasal lexicon is expected to be especially helpful to overcome some difficulties related to the context which cannot be handled well with standard dictionaries.

The phrasal lexicon used in our experiments consists of about ten thousand English phrases and their translations into Spanish and German. These English phrases have been extracted partly from various dialogue corpora and web-sites and have been partly created manually. The domain is not determined although a number of phrases contain dialogues. The average phrase length is about four words per entry. However, the vocabularies are rather large for all three languages containing as they do more than ten thousand distinct words. More than 60% of these words are singletons. Summarised statistics can be seen in Tables A.1 and A.2.

Short phrases: The short phrases[3] used as an additional bilingual knowledge source for Serbian–English translation contain about three hundred and fifty standard words and short expressions with an average entry length of 1.8 words for Serbian and two words for English. These phrases, although very scarce, might still be useful additional training material for a very small bilingual corpus. The statistics for these phrases are presented in Table A.4.

5.1.3 Small task-specific corpora

Spanish–English: The translation systems for this language pair are tested on the same TC-STAR corpora used in the experiments dealing with the local reorderings described in Section 4.3. In order to investigate sparse training data scenarios, two sets of small training corpora have been constructed by the random selection of sentences from the original corpus. The small corpus referred to as 13k contains about 1% of the original corpus. The corpus referred to as 1k contains only one thousand sentences; such a corpus can be produced manually in a relatively short time. In Table A.1 the statistics of these corpora can be found.

German–English: A small subset containing one thousand sentences is randomly extracted from the original EUROPARL training corpus. The translation is performed on the same test corpus as the long-range reordering experiments in Section 4.3. Corpus statistics are presented in Table A.2.

Serbian–English: The manually created electronic form of a language course contains two thousand and six hundred sentences and twenty five thousand running words of various types of conversations and descriptions as well as a few short newspaper articles. The average sentence

[3] `http://www.travlang.com/languages/ - Foreign languages for travellers`

length for Serbian is about 8.5 words, and for English about 9.5. Although the full corpus is already scarce, in order to investigate a scenario with extremely scarce training material, a reduced training corpus was created by random extraction of two hundred sentences from the original training corpus. The translation is performed on the test part of the language course corpus containing two hundred and sixty sentences. In order to examine the effects of translating data not related to a specific domain, we also translated a small test set collected from the BBC News web-site. Table A.4 contains the statistics for all corpora.

5.2 Morpho-syntactic transformations

Spanish–English

For this language pair, the local reorderings of nouns and adjectives described in Section 4.1 are applied for both translation directions. Apart from this, for translation into English, all Spanish adjectives are replaced with their base forms. The motivation for this is the following: Spanish adjectives, in contrast to English, have four possible inflectional forms depending on gender and number. This introduces additional data sparseness problems, especially if only a small amount of training data is available. It should be noted that although Spanish verbs have a much richer morphology than adjectives, we do not replace them with base forms: most verb inflections carry important information for translation into English, such as person, tense, etc.

German–English

For German–English translation with scarce resources, two types of morpho-syntactic transformations are applied: the long-range reordering of verbs as described in Section 4.2 and the splitting of compound words.

German compound words pose special problems to statistical machine translation systems: occurrences of each of the components in the training data is not sufficient for successful translation. Even if the compound itself has been seen in the training, the system may not be capable of translating it properly into two or more words. This can be particularly problematic if only a small amount of training material is available. We perform the frequency-based splitting proposed in [Koehn & Knight 03] in the following way:

- each capitalised word which consists of two or more words occuring in the training vocabulary is considered as a compound word;
- for each compound word:
  - the frequency of the compound itself $N(w)$ and the frequencies of its components $N(w_1),...,N(w_K)$ are collected
  - the geometric mean of the component frequencies is calculated $GM(f_1,...,f_K) = (\prod_{k=1}^{K} N(f_k))^{\frac{1}{K}}$
  - the compound word is split if $GM(f_1,...,f_K) > N(f)$

Serbian–English

The rich inflectional morphology of the Serbian language poses problems, especially for translation with scarce resources. The Serbian full forms of the words usually contain information which is not relevant for translation into English. Therefore we convert all Serbian words into their base forms. Nevertheless, the inflections of Serbian verbs might contain relevant information about person and tense, which is especially important if the pronoun is omitted (as in the case of Spanish verbs). Apart from this, there are three Serbian verbs which are negated by appending the negative particle to the verb as a prefix. Thus, the additional treatment of the Serbian verbs is applied. Whereas all other word classes are still replaced only by their base forms, for each verb a part of the POS tag referring to the person is taken and the verb is converted into a sequence of this tag and the base form. For the three verbs with a prefix negation, the separation of the negative particle from the verb is also applied so that each negative full form is transformed into the sequence of the POS tag, negative particle and base form. The transformed Serbian corpora contain significantly fewer singletons and OOV words than the original ones.

For translation from English into Serbian, we remove all articles from the English part of the corpus; articles are one of the most frequent word classes in English, but on the other hand there are no articles at all in Serbian. This method significantly reduces the number of running words and the average sentence length of the English corpus thus becoming more comparable to the corresponding values of the Serbian corpus.

5.3 Experimental results

Spanish–English language pair

The following set-ups are defined for the Spanish–English language pair:

- training only on the conventional dictionary (dictionary);
- training on the very small task-specific bilingual corpus (1k);
- training on the small task-specific bilingual corpus (13k);
- training on the large task-specific bilingual corpus (1.3M).

The language model for all set-ups is trained on the large corpus.

In this section, the results for the test corpus used in the second TC-STAR evaluation are presented. Results obtained on the test corpus from the first evaluation and on the Spanish Parliament test corpus can be found in Appendix C.

Spanish→English: Table 5.1 presents the results for translation from Spanish to English. As expected, the error rates of the system trained only on the dictionary are rather high, and morpho-syntactic transformations improve the performance. For this system, reorderings are applied only on the test corpus since the dictionary contains only short entries. Reducing Spanish adjectives to base forms decreases the OOV rate by 2% absolute and further improves the translation quality. An additional experiment with OOVs is performed for this set-up: all unseen full form words whose base form has been seen in the dictionary are replaced by this base form. This leads to a significant reduction of the OOV rate and further improvements of

Table 5.1: Translation results and OOV rates [%] for Spanish→English translation of the TC-STAR corpus: different sizes of training corpora and appropriate morpho-syntactic transformations.

Spanish→English	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	19.2	60.4	62.5	49.7	58.3	19.7
+reorder adjectives	21.5	58.8	60.7	49.4	56.4	19.7
+adjective base	22.8	57.6	59.5	48.0	55.2	17.7
+ OOV base	23.8	56.2	58.3	46.3	53.9	7.5
1k	27.8	52.2	54.1	41.8	50.2	11.9
+dictionary	34.0	47.2	49.6	36.9	46.4	7.3
+reorder adjectives	37.4	45.1	47.1	36.1	43.7	7.3
+adjective base	37.8	44.9	46.9	36.0	43.4	6.6
13k	41.6	41.6	44.2	31.6	40.6	3.8
+dictionary	43.2	40.3	43.0	31.0	40.2	2.9
+reorder adjectives	45.3	39.3	41.6	30.6	38.0	2.9
+adjective base	45.2	39.3	41.7	30.6	38.0	2.9
1.3M	52.0	34.6	36.9	26.3	33.7	0.6
+reorder adjectives	52.5	34.4	36.7	26.3	33.4	0.6

the translation performance. Although the final error rates after all morpho-syntactic transformations are still high, they might be acceptable for tasks where only the gist of the translated text is needed, like for example document classification or multilingual information retrieval. Additional morpho-syntactic transformations such as treatment of Spanish verbs could further improve the performance.

When only a very small amount of task-specific bilingual parallel text is used (1k), all error rates are decreased and the BLEU score is increased in comparison to the system trained on the dictionary alone, although they still remain rather high. Furthermore, it can be seen that the dictionary is very helpful as an additional training corpus and the morpho-syntactic transformations have a significant impact so that the final error rates are reduced by about 15% relative in comparison to the baseline system. By increasing the size of the task-specific training corpus (13k), all error rates decrease further and can be further reduced with help of the dictionary and morpho-syntactic transformations.

The best results obtained with the large corpus are about 12% relative better than the best results with the small corpus (13k) and about 25% relative better in comparison with the very small corpus (1k). These differences seem to be very large, but we have to keep in mind how large the differences between the corpus sizes are, especially in terms of the time and effort necessary for collecting and handling large corpora.

It should be noted that the impact of a dictionary as additional training material has not been tested for the full corpus since the corpus itself is sufficiently large. The replacement of Spanish adjectives with their base forms is also not tested on the full corpus since improvements are already insignificant for the 13k corpus. Apart from this, it can be seen that the local reorderings lead to more significant improvements for small training corpora. This happens because the baseline phrase-based translation system can handle very well local word order differences

which are seen sufficiently often in the training corpus. However, the unseen or rarely seen phrases pose reordering problems, and in small corpora the number of such phrases is much higher.

An illustration of translation with scarce resources is presented in Table 5.2. With the full training corpus, the system produces a completely correct output. The system trained on the small corpus (13k) fails to produce the correct word order, and the problem is solved by applying the local adjective reorderings. When only the very small training corpus is used (1k), the translation system has difficulties both with word order and with the unseen adjective form "distinta". By applying local word reorderings and reduction of adjectives into bases, the system becomes capable to produce the correct output. Generating correct output becomes even more difficult when only the dictionary is used as a training corpus: apart from the wrong word order, two unseen words ("tienen" and "distinta") and one extra word ("states") is present. Reordering of the test corpus and reducing all adjectives and unseen full forms to base forms significantly improves the translation output, but it is still not correct because the extra word remains.

Table 5.2: Example of Spanish→English translation with different sizes of training corpora with and without transformations.

Spanish sentence:	Los jóvenes tienen una visión distinta de Europa.
generated English sentence 1.3M:	The young people have a different vision of Europe.
13k:	The young people have a *vision different* of Europe.
+reorder:	The young people have a different vision of Europe.
1k:	The young people have a *vision distinta* of Europe.
+reorder+adj-bases:	The young people have a different vision of Europe.
dictionary:	*States* young people *tienen* a *vision distinta* of Europe.
+reorder+OOV-bases:	*States* young people have a different vision of Europe.
reference English sentence:	The young people have a different vision of Europe.

English→Spanish: The translation results for the other direction can be seen in Table 5.3. Error rates are higher due to the inflectional morphology of the Spanish language, and the effects of the training corpus size, dictionary and morpho-syntactic transformations are very similar as for the translation into English. The improvements from the morpho-syntactic transformations are slightly smaller than for the translation into English due to the phenomenon already described in previous sections; in the Spanish language the adjective group is not always situated behind the noun. Nevertheless, for this translation direction the same phenomenon can be noted; local reorderings are especially important for the small training corpora.

Translation of word graphs

This section presents results of the translation of word graphs as described in Section 4.1 by systems trained on the sparse training corpora. Like in the experiments described in Section 4.3, training is done both with the original corpus (baseline) and with the reordered source language corpus (reorder adjectives) except for the dictionary as explained in the previous section.

Table 5.3: Translation results and OOV rates [%] for English→Spanish translation of the TC-STAR corpus: different sizes of training corpora and appropriate morpho-syntactic transformations.

English→Spanish	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	15.6	65.3	66.7	55.8	63.6	14.9
+reorder adjectives	17.5	63.8	65.2	54.8	62.0	14.9
1k	22.6	59.1	60.8	49.2	57.1	10.5
+dictionary	25.8	55.3	57.2	45.8	53.8	4.9
+reorder adjectives	27.5	54.9	56.8	44.9	53.2	4.9
13k	36.7	47.0	49.1	37.8	46.0	2.8
+dictionary	37.6	46.4	48.4	37.2	45.6	2.0
+reorder adjectives	39.4	44.8	46.8	36.4	44.1	2.0
1.3M	48.2	38.5	40.4	31.1	37.6	0.4
+reorder adjectives	49.0	38.1	39.8	30.8	37.0	0.4

Spanish→English: The translation results are presented in Table 5.4. For the dictionary, the use of the graph significantly improves the translation performance, but the best results are obtained by the fixed reordering of the test set. Similar tendencies can be seen for both small task-specific corpora.

Table 5.4: Translation results and OOV rates [%] on the small training corpora for the Spanish→English reordering word graphs.

Spanish→English	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	19.2	60.4	62.5	49.7	58.3	19.7
+graph	**21.5**	59.0	60.8	**49.4**	56.6	
+reorder adjectives	**21.5**	**58.8**	**60.7**	**49.4**	**56.4**	
1k	27.8	52.2	54.1	41.8	50.2	11.9
+graph	29.8	50.7	52.6	41.4	48.8	
reorder adjectives	**30.1**	**50.2**	**52.1**	**40.8**	**48.2**	
+graph	**30.1**	50.3	52.2	**40.8**	48.3	
13k	41.6	41.6	44.2	31.6	40.6	3.8
+graph	**44.1**	**40.2**	42.5	**31.3**	**38.8**	
reorder adjectives	43.9	**40.2**	**42.4**	31.5	**38.8**	
+graph	44.0	**40.2**	42.5	31.4	38.9	

English→Spanish: As for the large training corpus, the graphs seem to be more important for this translation direction. Table 5.5 shows that for the training on the dictionary alone, the translation of word graphs yields the best translation performance. For the very small corpus (1k), the best performance is achieved with reordering in training and word graph translation, whereas for the small corpus (13k) word graph translation using the baseline system outperforms other configurations. However, the results for the reordered training corpus and the word graph are very close to the best ones.

Table 5.5: Translation results and OOV rates [%] on the small training corpora for the English→Spanish reordering word graphs.

English→Spanish	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	15.6	65.3	66.7	55.8	63.6	14.9
+graph	**18.0**	**63.5**	**64.8**	**54.7**	**61.6**	
+reorder adjectives	17.5	63.8	65.2	54.8	62.0	
1k	22.6	59.1	60.8	49.2	57.1	10.5
+graph	24.4	57.3	**58.9**	48.2	55.4	
reorder adjectives	24.0	57.6	59.5	47.8	55.5	
+graph	**24.7**	**57.1**	**58.9**	**47.6**	**55.0**	
13k	36.7	47.0	49.1	37.8	46.0	2.8
+graph	**39.4**	**45.1**	**47.0**	**36.8**	**44.0**	
reorder adjectives	38.2	45.6	47.6	37.1	44.6	
+graph	39.1	45.3	47.1	37.0	44.1	

Phrasal lexicon

This section presents translation results for the Spanish–English phrasal lexicon described in Section 5.1.2. For exploring the translation with phrasal lexicon, the following set-ups are defined:

- training on the phrasal lexicon (phrases);
- training on the reordered phrasal lexicon (reorder adjective);
- training on the phrasal lexicon together with the conventional dictionary (phrases+dictionary).

The language model is trained on the large task-specific corpus, as for the experiments described in the sections above.

Translation results are presented in Tables 5.6 and 5.7. For both translation directions, the best configuration is translation of word graph without reordering in training. For Spanish→English translation, these results are very similar to those obtained with fixed reorderings in training and test. For the other translation direction this best configuration clearly outperforms all other.

Table 5.6: Translation results and OOV rates [%] on the phrasal lexicon for Spanish→English: local reorderings and word graphs.

Spanish→English	BLEU	TER	WER	PER	CDER	OOV rate
phrases	22.8	56.5	58.4	46.6	54.4	22.4
+graph	**24.4**	55.3	**57.0**	**46.3**	**53.1**	
reorder adjectives	24.2	**55.2**	**57.0**	46.4	**53.1**	
+graph	24.1	55.5	57.3	46.4	53.4	
phrases+dictionary	29.4	51.5	53.6	41.2	49.7	14.2
+graph	32.0	49.8	51.8	40.7	47.9	

Table 5.7: Translation results and OOV rates [%] on the phrasal lexicon for English→Spanish: local reorderings and word graphs.

English→Spanish	BLEU	TER	WER	PER	CDER	OOV rate
phrases	17.8	63.2	64.7	53.4	61.4	17.1
+graph	**19.4**	**61.8**	**63.1**	**52.8**	**59.9**	
reorder adjectives	18.4	62.4	63.8	53.3	60.6	
+graph	19.2	62.0	63.3	53.0	60.2	
phrases+dictionary	23.1	58.4	60.1	48.6	56.8	9.9
+graph	25.8	56.4	57.9	47.6	54.8	

Joining the phrasal lexicon with the conventional dictionary reduces the number of OOVs, and the translation results significantly outperform those obtained on one of the corpora alone. Using word graphs for such a translation system further improves the translation quality for both translation directions.

German–English language pair

For the German–English pair the following configurations are defined:

- training only on the conventional dictionary (dictionary);
- training on the very small task-specific bilingual corpus (1k);
- training on the large task-specific bilingual corpus (700k).

The language model for all systems is again trained on the large corpus.

German→English: Results are presented in Table 5.8. The performance of the dictionary alone is similar to the performance of the 1k corpus; the OOV rate for the small corpus is much higher since the dictionary has a very rich vocabulary as described in Section 5.1.1. However, the sentences in the dictionary are short so that syntactic phenomena are handled better with the task-specific corpus. Long-range verb reorderings improve translation performance for all setups, and additional gains are obtained by compound splitting.[4] The improvements obtained by verb reorderings are similar for all corpora, and the compound splitting has a larger impact for the small training sets. More results of compound splitting for the large corpus can be found in Section C.1 in Appendix C. The best results on the large corpora are about 15% relative better than the best results with the small corpus (0.14% of the full corpus size) with the dictionary.

English→German: Table 5.9 shows the results for translation into German. Again, error rates for the dictionary are similar to those with the 1k corpus and are improved by long-range infinitive and past participle reorderings. However, it can be observed that for this translation direction, improvements from the reorderings are smaller for the small corpus than for the large one. The most probable reason for this is precisely the long-range, as already mentioned in Section 4.3; long distances between words within a phrase generally pose problems for translation

[4]The approach used in this work is corpus-based which could raise the question as to whether splitting can be done adequately if only a small amount of text is available. However, we follow the same reasoning as for language model training; since for the splitting only a monolingual German corpus is needed, we perform all splittings learnt from the largest corpus.

systems by generating rare patterns even if the word order in the source and target language is harmonised. These problems are even greater if only a small training corpus is available.

Table 5.8: Translation results and OOV rates [%] for German→English translation of the EUROPARL corpus: different sizes of training corpora and appropriate morpho-syntactic transformations.

German→English	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	11.7	74.8	78.5	59.2	67.8	10.2
+reorder verbs	12.5	73.4	77.1	58.6	66.6	10.2
+split compounds	12.8	73.0	76.8	57.9	66.2	9.6
1k	11.6	75.2	78.5	60.3	68.5	16.4
+dictionary	14.6	71.6	75.8	55.9	64.8	6.6
+reorder verbs	15.0	70.7	74.9	55.4	64.0	6.6
+split compounds	15.7	70.0	74.2	54.5	63.4	5.6
700k	24.4	61.3	66.9	45.9	55.6	0.8
+reorder verbs	25.6	60.2	65.4	45.7	54.4	0.8
+split compounds	25.6	59.8	65.1	45.2	54.4	0.7

Table 5.9: Translation results and OOV rates [%] for English→German translation of the EUROPARL corpus: different sizes of training corpora and appropriate morpho-syntactic transformations.

English→German	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	8.9	80.4	83.1	65.7	71.4	4.2
+reorder verbs	9.3	80.0	82.8	65.6	70.8	4.2
1k	8.4	82.9	85.5	68.1	72.1	10.2
+dictionary	10.8	78.9	81.9	64.0	69.1	2.2
+reorder verbs	11.0	78.7	81.8	64.0	68.9	2.2
700k	18.2	68.6	72.6	54.2	61.3	0.2
+reorder verbs	18.4	67.8	71.6	53.8	61.0	0.2

Translation of word graphs

The results of the translation of the German word graphs described in Section 4.2 using sparse training corpora are shown in Table 5.10. Training is again done with the original corpus (baseline) as well as with the reordered source language corpus (reorder verbs). Reordering in training is examined also for the dictionary because it contains a number of short sentences, in contrast to the Spanish–English dictionary.

Similar to the case of the full training corpus (Section 4.3), the best solution is to translate word graphs with additional local reorderings. However, it can be noted that for the dictionary, the best performance is achieved when the reordering in training is applied, whereas for the 1k the better solution is to translate with the original system. The same tendencies are observed for

Table 5.10: Translation results and OOV rates [%] on the small training corpora for the German→English reordering word graphs.

German→English	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	11.7	74.8	78.5	59.2	67.8	10.2
+graph	12.6	73.5	77.2	58.7	66.6	
+local	12.6	73.7	77.1	58.6	66.5	
reorder verbs	12.6	73.4	77.1	58.6	66.6	
+graph	12.8	73.2	76.9	58.5	**66.4**	
+local	**12.9**	**73.1**	**76.8**	**58.4**	**66.4**	
1k	11.6	75.2	78.5	60.3	68.5	16.4
+graph	12.2	**74.4**	**77.7**	**60.0**	67.7	
+local	**12.4**	**74.4**	77.8	**60.0**	**67.6**	
reorder verbs	11.9	74.8	78.2	60.2	67.9	
+graph	12.1	74.7	78.0	60.2	67.8	
+local	12.3	74.8	78.2	60.4	**67.6**	

Table 5.11: Translation results and OOV rates [%] on the small training corpora for the English→German reordering word graphs.

English→German	BLEU	TER	WER	PER	CDER	OOV rate
dictionary	8.9	80.4	83.1	65.7	71.4	4.2
+graph	9.1	80.1	82.8	65.6	71.1	
reorder verbs	9.3	80.1	82.8	65.6	70.8	
+graph	**9.4**	**79.9**	**82.6**	**65.3**	**70.7**	
1k	8.4	82.9	85.5	68.1	72.1	10.2
+graph	**8.5**	**82.5**	**85.0**	**68.0**	71.8	
reorder verbs	8.4	83.2	85.7	68.9	**71.7**	
+graph	8.4	83.4	85.9	68.9	**71.7**	

the other translation direction presented in Table 5.11 which can be explained by phenomena already mentioned in previous sections, i.e. that positioning of verbs away from their corresponding pronouns/auxiliaries/modals introduces certain difficulties for statistical models, especially for small training corpora. However, for translation from German into English a straighforward explanation for this phenomenon is not yet easy to find. In addition, all the results are quite close. More experiments and detailed error analysis should be performed before making any conclusions.

Phrasal lexicon

Analogously to the Spanish–English language pair, the following set-ups are defined for translation using the German–English phrasal lexicon:

- training on the phrasal lexicon (phrases);

- training on the reordered phrasal lexicon (reorder verbs);
- training on the phrasal lexicon together with the conventional dictionary (phrases+dictionary).

Translation from German into English using the phrasal lexicon (Table 5.12) yields the best results with the word graph and reorderings in training. The same configuration produces the best results also when both phrasal lexicon and dictionary are used for training, although very similar performance is obtained without reorderings in training. For translation into German (Table 5.13), the best option is to translate word graphs without reorderings in training.

Table 5.12: Translation results and OOV rates [%] on the phrasal lexicon for German→English: long-range reorderings and word graphs.

German→English	BLEU	TER	WER	PER	CDER	OOV rate
phrases	6.5	80.7	83.5	66.0	73.2	25.4
+graph	6.9	80.0	82.8	65.8	72.4	
+local	7.1	79.6	82.4	65.6	**72.1**	
reorder verbs	7.0	78.7	81.4	64.8	72.5	
+graph	7.1	78.9	81.6	64.8	72.6	
+local	**7.3**	**78.3**	**81.0**	**64.5**	**72.1**	
phrases+dictionary	12.8	73.5	77.4	58.1	66.7	9.1
+graph	13.5	72.4	76.2	57.6	65.6	
+local	**13.6**	72.2	76.1	57.6	65.6	
reorder verbs	13.2	72.6	76.4	57.7	65.8	
+graph	13.4	72.3	76.2	**57.4**	65.6	
+local	13.5	**72.1**	**76.0**	**57.4**	**65.4**	

Table 5.13: Translation results and OOV rates [%] on the phrasal lexicon for English→German: long-range reorderings and word graphs.

English→German	BLEU	TER	WER	PER	CDER	OOV rate
phrases	5.2	87.5	89.7	73.4	76.5	17.8
+graph	**5.3**	**87.0**	89.2	**73.2**	**76.3**	
reorder verbs	5.2	87.1	**89.1**	**73.2**	76.4	
+graph	5.2	**87.0**	**89.1**	**73.2**	76.4	
phrases+dictionary	9.5	79.7	82.6	64.9	70.5	3.9
+graph	9.6	**79.4**	**82.2**	**64.8**	70.4	
reorder verbs	**9.7**	79.7	82.4	65.2	**70.3**	
+graph	**9.7**	79.6	82.4	65.0	**70.3**	

Serbian–English language pair

For this language pair the following training set-ups are defined:

- training on an extremely small task-specific bilingual corpus (0.2k);
- training on a small task-specific bilingual corpus (2.6k).

Since the largest available corpus is already small and the external phrase book is even smaller, we have not investigated translation using only the phrase book, but we used it as additional training material for the extremely sparse training corpus. The language model for all set-ups was trained on the full (2.6k) corpus.

Error rates for translation from Serbian into English are shown in Table 5.14. As expected, the error rates of the system trained on an extremely small amount of parallel data are very high. Performance of such a system is comparable with a system trained only on a conventional dictionary. Adding short phrases is helpful to some extent, and replacing words with base forms has the largest impact by almost halving the OOV rate and decreasing all error rates significantly. Further improvements in PER, CDER and the BLEU score are obtained by the verb treatment described in Section 5.2, although TER and WER are slightly deteriorated. Increasing the size of the bilingual training corpus to about three thousand sentences and applying morpho-syntactic transformations leads to an improvement of about 30% relative. Using a conventional dictionary and additional morpho-syntactic transformations should further improve the performance.

Table 5.14: Translation results and OOV rates [%] for Serbian→English translation: different sizes of training corpora and appropriate morpho-syntactic transformations.

Serbian→English	BLEU	TER	WER	PER	CDER	OOV rate
0.2k	8.3	65.0	65.5	60.8	63.3	35.2
+phrases	10.3	64.3	65.0	59.8	62.1	31.8
+base forms	13.9	**58.4**	**59.2**	54.8	57.2	19.3
+verb treatment	**14.8**	58.9	60.0	**52.6**	**55.4**	16.3
2.6k	32.1	43.1	44.5	37.9	41.4	11.7
+base forms	35.4	41.8	42.9	37.4	39.8	4.7
+verb treatment	**36.4**	**40.4**	**41.9**	**34.7**	**38.2**	3.9

From the first Serbian-to-English translation example in Table 5.15, it can be seen how the problem of some OOV words can be overcome with the use of the base forms. The second and third examples show the advantages of the verb treatment: the second one presents the introduction of relevant parts of the verb POS tags and the third one illustrates the effect of separating the negative particle.

Table 5.16 shows results for translation from English into Serbian. As expected, all error rates are significantly higher than for the other translation direction since the translation into the morphologically richer language always has poorer quality. The importance of the phrases seems to be larger for this translation direction. Removing the English articles improves the translation quality for both set-ups. As in the case of the other translation direction, increasing the size of the training corpus results in up to 30% relative improvement.

Table 5.15: Examples of Serbian→English translations with and without moprho-syntactic transformations.

original Serbian sentence:	to *je* suviše *skupo*.
base forms:	to *biti* suviše *skup*.
+verb treatment:	to *SG3 biti* suviše *skup*.
generated English sentence	
without transformations:	it is too **UNKNOWN_skupo**.
with base forms:	it is too **expensive**.
with verb treatment:	it is too expensive.
reference English sentence:	it is too expensive.
original Serbian sentence:	on ne *igra*.
base forms:	on ne *igrati*.
+verb treatment:	on ne *SG3 igrati*.
generated English sentence	
without transformations:	he **he** does not.
with base forms:	he **do** not play.
with verb treatment:	he **does** not play.
reference English sentence:	he does not play.
original Serbian sentence:	da, ali *nemam* mnogo vremena.
base forms:	da, ali *nemati* mnogo vreme.
+verb treatment:	da, ali *SG1 ne imati* mnogo vreme.
generated English sentence	
without transformations:	yes, but **I have** much time.
with base forms:	yes, but **not** much time.
with verb treatment:	yes, but **I have not got** much time.
reference English sentence:	yes, but I have not much time.

Table 5.16: Translation results and OOV rates [%] for English→Serbian translation: different sizes of training corpora and appropriate morpho-syntactic transformations.

English→Serbian	BLEU	TER	WER	PER	CDER	OOV rate
0.2k	6.8	72.9	73.4	68.4	65.7	21.8
+phrases	9.3	71.5	71.9	67.5	64.6	18.8
+remove articles	9.4	66.4	66.7	62.2	62.3	20.0
2.6k	23.1	51.1	51.8	45.8	48.7	4.9
+remove articles	24.6	49.6	50.4	44.6	47.3	5.3

Results for the out-of-domain BBC text are shown in Table 5.17 for both translation directions. The number of OOV words and the error rates are very high and can be compared with a system trained on a conventional dictionary. A significant decrease in the number of OOV words when translating from Serbian is achieved by the use of morpho-syntactic transformations. The other translation direction is slightly improved by removing the English articles.

Table 5.17: Translation results and OOV rates [%] for the out-of-domain BBC News text: training on the 2.6k corpus with and without morpho-syntactic transformations.

		BLEU	TER	WER	PER	CDER	OOV rate
Serbian→English	2.6k	9.8	70.2	70.6	65.2	69.3	44.3
	+base forms	13.6	66.8	67.0	60.5	64.3	35.4
	+verb treatment	14.6	66.1	66.8	59.0	64.3	31.3
English→Serbian	2.6k	5.0	77.2	78.5	71.9	71.9	32.1
	+remove articles	5.2	75.9	77.0	71.1	71.4	34.7

5.4 Conclusions

A thorough investigation of translation with sparse bilingual resources was carried out, and systematic experiments on three distinct language pairs and different types of corpora have shown that an acceptable translation quality can be achieved with a very small amount of task-specific parallel text, especially if conventional dictionaries, phrasal books, as well as morpho-syntactic knowledge are available. Translation systems built only on a conventional dictionary, phrasal lexicon or on extremely small task-specific corpora might be useful for applications such as document classification or multilingual information retrieval. With the help of dictionaries/lexica and proper morpho-syntactic transformations, an acceptable translation quality can be achieved with only one thousand sentence pairs of in-domain text. The big advantage of such a small corpus is that the costs of its acquisition are rather low; such a corpus can be manually produced in a relatively short time.

The particular effects related to language characteristics, type and size of the corpus and applied morpho-syntactic transformations are also studied, and the following phenomena are observed:

- local reorderings of adjectives are very helpful for the small training corpora, much more than for the large corpora;
- long-range reorderings in German improve the translation quality for all corpora;
- long-range reorderings in English work better on a large corpus;
- translation of word graphs with small training corpora results in the same tendencies as for the large corpora; however, the German→English translation direction remains an exception – more experiments as well as a detailed analysis are needed;
- reducing Spanish adjectives to base forms improves the performance for the very small corpora and dictionaries;
- German compound splitting seems to be more important for the small corpora;
- morpho-syntactic treatment of (highly inflected) Serbian words significantly improves translation into English;
- it would be interesting to compare the results for the Serbian–English language pair with the results obtained on large corpora and on conventional dictionaries/phrasal lexica.

6 Automatic error analysis of translation output

Evaluation and error analysis of machine translation output are very important tasks, but difficult both for machines and humans. Human evaluation is expensive and time-consuming. Whereas many automatic evaluation measures are available, and some of them are widely used, automatic error analysis of translation output is mostly an unexplored area.

In this work, a framework for automatic error analysis and categorisation is presented. The basic idea is to actually identify erroneous words using the algorithms for the calculation of WER and PER. The extracted error details can be used in combination with different types of linguistic knowledge (such as base forms, POS tags, NE tags, compound words, suffix, prefix, etc.) in order to obtain various details about actual errors (for example error categories (e.g. morphological errors, reordering errors, missing words), contribution of different word classes (e.g. POS tags, NE tags), etc.).

The following error categories are in the focus of this work: morphological (inflectional) errors, reordering errors, missing words, extra words and incorrect lexical choice. Each error category can be futher classified according to POS tags (e.g. inflectional errors of verbs, missing pronouns, etc.).

In this thesis, a comparison of the results of automatic error analysis with those obtained by human error analysis for these error categories is carried out. Furthermore, new error rates based on the proposed error categories are introduced and used for the comparison of different translation systems. We examine how the changes within one translation system, as well as the differences between translation systems, are reflected in the new measures. In addition, an alternative method for automatic estimation of reordering and inflectional errors is proposed.

6.1 Framework for automatic error analysis

The basic idea for automatic error analysis is to take details from WER (edit distance) and PER algorithms, namely to identify all words which actually contribute to the error rate, and then to combine different types of linguistic knowledge of these words. The general procedure of automatic error analysis and classification is shown in Figure 6.1. An overview of the standard word error rates WER and PER is given in Section 6.1.1, and methods for extracting actual errors are described in the following sections.

In this thesis, we carried out the error analysis on the word level and we used base forms of the words and POS tags as linguistic knowledge. However, the analysis presented in this work is only one of many possibilities – this framework enables the integration of various

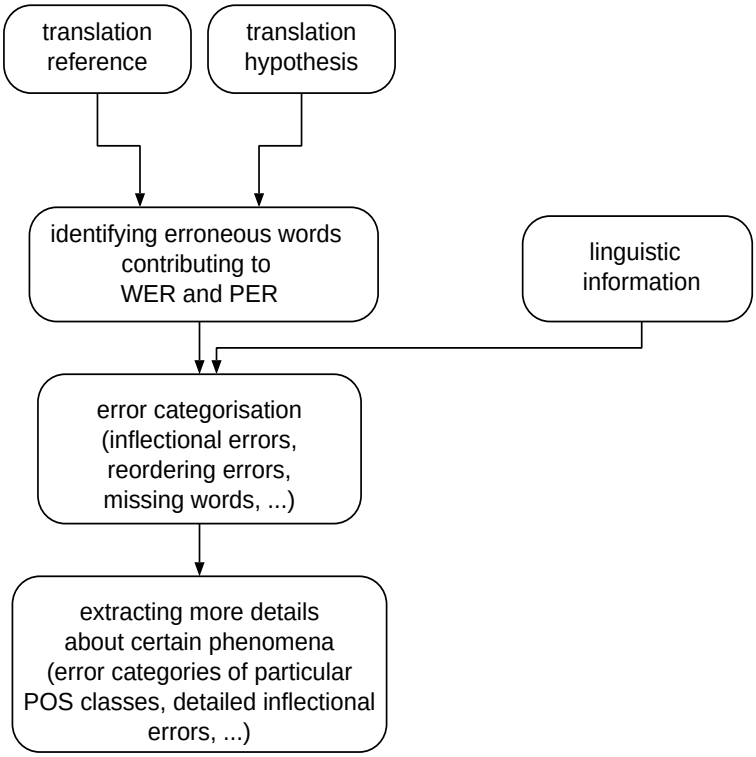

Figure 6.1: General procedure for automatic error analysis based on the standard word error rates and linguistic information.

knowledge sources such as deeper lingustic knowledge, introduction of source words (possibly with additional linguistic information) if appropriate alignment information is available, etc. Investigation on the word group/phrase level instead of only on the word level is possible as well.

6.1.1 Standard word error rates (overview)

The standard procedure for evaluating machine translation output is done by comparing the hypothesis document *hyp* with the given reference document *ref*, each one consisting of K sentences (or segments). The reference document *ref* consists of $N_R \geq 1$ reference translations of the source text. $N_R = 1$ stands for the case when only a single reference translation is available, and $N_R > 1$ denotes the case of multiple references. Let the length of the hypothesis sentence hyp_k be denoted as N_{hyp_k}, and the length of each reference sentence $N_{ref_{k,r}}$. Then, the total hypothesis length of the document is $N_{hyp} = \sum_k N_{hyp_k}$ and the total reference length is $N_{ref} = \sum_k N^*_{ref_k}$ where $N^*_{ref_k}$ is defined as the length of the reference sentence with the lowest sentence–level error rate as shown to be optimal with respect to the correlation with the human evaluation scores adequacy and fluency [Leusch & Ueffing$^+$ 05]. The overall error rate is then obtained by normalising the total number of errors over the total reference length.

The word error rate (WER) is based on the Levenshtein distance [Levenshtein 66] – the minimum number of substitutions, deletions and insertions that have to be performed to convert the

generated text *hyp* into the reference text *ref*. A shortcoming of the WER is the fact that it does not allow reorderings of words, whereas the word order of the hypothesis can be different from the word order of the reference even though it is a correct translation. The position-independent word error rate (PER) is also based on substitutions, deletions and insertions but without taking the word order into account. The PER is always lower than or equal to the WER. On the other hand, a shortcoming of the PER is the fact that it does not penalise a wrong word order.

Calculation of WER: The WER of the hypothesis *hyp* with respect to the reference *ref* is calculated as:

$$\text{WER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \min_{r} d_L(ref_{k,r}, hyp_k)$$

where $d_L(ref_{k,r}, hyp_k)$ is the Levenshtein distance between the reference sentence $ref_{k,r}$ and the hypothesis sentence hyp_k. The calculation of WER is performed using a dynamic programming algorithm.

Calculation of PER: Define $n(w, setw)$ as the number of occurrences of a word w in a set of words $setw$. The PER can be calculated using the counts $n(e, hyp_k)$ and $n(e, ref_{k,r})$ of a word e in the hypothesis sentence hyp_k and the reference sentence $ref_{k,r}$ respectively:

$$\text{PER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \min_{r} d_{\text{PER}}(ref_{k,r}, hyp_k)$$

where

$$d_{\text{PER}}(ref_{k,r}, hyp_k) = \frac{1}{2}\left(|N_{ref_{k,r}} - N_{hyp_k}| + \sum_{e} |n(e, ref_{k,r}) - n(e, hyp_k)| \right)$$

6.1.2 Identification of WER errors

The dynamic programming algorithm for WER enables a simple and straightforward identification of each erroneous word which actually contributes to WER. Let err_k denote the set of erroneous words in sentence k with respect to the best reference and e be a word. Then $N(wer) = n(e, err_k)$ is the number of WER errors in err_k.

An example of a reference sentence and hypothesis sentence is shown in Table 6.1. The WER errors, i.e. actual words participating in WER can be seen in Table 6.2. The reference words involved in WER are denoted as reference errors, and hypothesis errors refer to the hypothesis words participating in WER. Table 6.3 presents an example of introducing linguistic knowledge, i.e. POS tags. The WER errors are identified along with their corresponding POS tags, and the contribution of each POS class to the overall WER is calculated.

If err_k is the set of erroneous words in sentence k with respect to the best reference and p is a POS class, then $N(wer(p)) = \sum_{e \in p} n(e, werr_k)$ is the number of WER errors in err_k produced by words belonging to the POS class p. It should be noted that for the substitution errors, the POS class of the involved reference word is taken into account. POS tags of the reference words are

also used for the deletion errors, and for the insertion errors the POS class of the hypothesis word is used. The WER for the word class p can be calculated like the standard WER by normalising the number of errors over the reference length:

$$\text{WER}(p) = \frac{1}{N_{ref}} \sum_{k=1}^{K} \sum_{e \in p} n(e, werr_k)$$

Table 6.1: Example for illustration of actual errors: a reference sentence and a corresponding hypothesis sentence.

reference:	hypothesis:
Mister Commissioner , twenty-four hours sometimes can be too much time .	Mrs Commissioner , sometimes twenty-four hours is too much time .

Table 6.2: WER errors: actual words which participate in the word error rate.

reference errors	hypothesis errors	error type
Mister	Mrs	substitution
	sometimes	insertion
sometimes	is	substitution
can		deletion
be		deletion

Table 6.3: WER errors and linguistic knowledge: actual words which are participating in the word error rate with their corresponding base forms and POS classes.

reference errors	hypothesis errors	error type
Mister#Mister#N	Mrs#Mrs#N	substitution
	sometimes#sometimes#ADV	insertion
sometimes#sometimes#ADV	is#be#V	substitution
can#can#V		deletion
be#be#V		deletion

The standard WER of the whole sentence in Table 6.3 is equal to $5/12 = 41.7\%$. The contribution of nouns is $\text{WER}(\text{N}) = 1/12 = 8.3\%$, of verbs $\text{WER}(\text{V}) = 2/12 = 16.7\%$ and of adverbs $\text{WER}(\text{ADV}) = 2/12 = 16.7\%$.

6.1.3 Identification of PER errors

In contrast to WER, the standard efficient algorithms for calculation of PER do not give precise information about contributing words. However, it is possible to identify all words in

the hypothesis which do not have a counterpart in the reference, and vice versa. These words will be referred to as PER errors.

An illustration of PER errors is given in Table 6.4. The number of errors contributing to the standard PER according to the algorithm described in Section 6.1.1 is 3 – there are two substitutions and one deletion. The problem with standard PER is that it is not possible to detect which words are the deletion errors, which are the insertion errors, and which words are the substitution errors. We introduce alternative PER-based measures which correspond to the precision, recall and F-measure. Let $herr_k$ refer to the set of words in the hypothesis sentence k which do not appear in the reference sentence k (referred to as hypothesis errors). Analogously, let $rerr_k$ denote the set of words in the reference sentence k which do not appear in the hypothesis sentence k (referred to as reference errors). Then the following measures can be calculated:

- recall-based (reference) PER (RPER):

$$\text{RPER} = \frac{1}{N_{ref}} \sum_{k=1}^{K} \sum_{e} n(e, rerr_k)$$

- precision-based (hypothesis) PER (HPER):

$$\text{HPER} = \frac{1}{N_{hyp}} \sum_{k=1}^{K} \sum_{e} n(e, herr_k)$$

- F-based PER (FPER):

$$\text{FPER} = \frac{1}{N_{ref} + N_{hyp}} \cdot \sum_{k=1}^{K} \sum_{e} \Big(n(e, rerr_k) + n(e, herr_k) \Big)$$

For the example sentence presented in Table 6.1, the number of hypothesis errors $n(e, herr_k)$ is 2 and the number of reference errors $n(e, rerr_k)$ is 3 where e denotes the word. The number of errors contributing to the standard PER is 3, since $|N_{ref} - N_{hyp}| = 1$ and $\sum_e |n(e, ref_k) - n(e, hyp_k)| = 5$. The standard PER is normalised over the reference length $N_{ref} = 12$ thus being equal to 25%. The RPER considers only the reference errors, $\text{RPER} = 3/12 = 25\%$, and HPER only the hypothesis errors, $\text{HPER} = 2/11 = 18.2\%$. The FPER is the sum of hypothesis and reference errors divided by the sum of hypothesis and reference length: $\text{FPER} = (2+3)/(11+12) = 5/23 = 21.7\%$.

Table 6.4: PER errors: actual words which participate in the position-independent word error rate.

reference errors	hypothesis errors
Mister	Mrs
can	is
be	

For the example of PER errors with corresponding POS tags in Table 6.5, contributions of nouns are $\text{RPER}(\text{N}) = 1/12 = 8.3\%$, $\text{HPER}(\text{N}) = 1/11 = 9.1\%$ and $\text{FPER}(\text{N}) = 2/23 = 8.7\%$, and the contributions of verbs $\text{RPER}(\text{V}) = 2/12 = 16.7\%$, $\text{HPER}(\text{V}) = 1/11 = 9.1\%$ and $\text{FPER}(\text{V}) = 3/23 = 13\%$.

Table 6.5: PER errors and linguistic knowledge: actual words which participate in the position-independent word error rate and their corresponding base forms and POS classes.

reference errors	hypothesis errors
Mister#Mister#N	Mrs#Mrs#N
be#be#V	is#be#V
can#can#V	

6.2 Methods for automatic error analysis and classification

The error details described in Section 6.1.2 and Section 6.1.3 can be combined with different types of linguistic knowledge in different ways. Examples with base forms and POS tags as linguistic knowledge are presented in Table 6.3 and 6.5. The described error rates of the particular POS classes give more details than the overall standard error rates and can be used for error analysis to some extent. However, for more precise information about certain phenomena some kind of further analysis is required. In this work, we examine the following error categories:

- inflectional errors – using PER errors and base forms;
- reordering errors – using WER and PER errors;
- missing words – using WER and PER reference errors with base forms;
- extra words – using WER and PER hypothesis errors with base forms;
- incorrect lexical choice – reference errors which belong neither to inflectional errors nor to missing words.

Furthermore, the contribution of the various POS classes for the described error categories is estimated.

It should be noted that the base forms and POS tags are needed both for the reference(s) and for the hypothesis. The performance of morpho-syntactic analysis is slightly lower on the hypothesis, but this does not seem to influence the performance of the error analysis tools. However, we choose to use reference words for all cases where it can be chosen between the reference and the hypothesis. Nevertheless, it would be interesting to investigate the use of hypothesis words in future experiments and compare the results.

Inflectional errors

An inflectional error occurs if the base form of the generated word is correct but the full form is wrong. Inflectional errors can be estimated using RPER errors and base forms in the following way: from each reference–hypothesis sentence pair, only erroneous words which have the common base forms are taken into account:

$$N(infl) = \sum_{k=1}^{K} \sum_{e} n(e, rerr_k) - \sum_{k=1}^{K} \sum_{eb} n(eb, rberr_k)$$

where *eb* denotes the base form of the word and $rberr_k$ stands for the set of base form errors in the reference. The number of words with erroneous base forms (representing non-inflectional errors) is substracted from the number of total errors. An analogous definition is possible using HPER errors and base forms – however, as explained at the beginning of this section, we choose to use the reference words because the results of the morpho-syntactic analysis are slightly more reliable for the references than for the hypotheses.

For example, from the PER errors presented in Table 6.4, the word "is" will be detected as an inflectional error because it shares the same base form with the reference error "be".

Reordering errors

The differences in word order in the hypothesis with respect to the reference are taken into account only by WER and not by PER. Therefore, a word which occurs both in the reference and in the hypothesis but is marked as a WER error is considered as a reordering error. The contribution of reordering errors can be estimated in the following way:

$$N(reord) = \sum_{k=1}^{K} \sum_{e} \Big(n(e, suberr_k) + n(e, delerr_k) - n(e, rerr_k) \Big)$$

where $suberr_k$ represents the set of WER substitution errors, $delerr_k$ the set of WER deletion errors and $rerr_k$ the set of RPER errors. For the example in Table 6.1, the word "sometimes" is identified as a reordering error.

Missing words

Missing words can be identified using the WER and PER errors in the following way; the words considered as missing are those which occur as deletions in WER errors and at the same time occur only as reference PER errors without sharing the base form with any hypothesis error i.e. as a non-inflectional RPER error:

$$N(miss) = \sum_{k=1}^{K} \sum_{eb \in rberr_k} n(e, delerr_k)$$

The set of deletion WER errors is defined as $delerr_k$, whereas $rberr_k$ stands for the set of base form RPER errors.

The use of both WER and PER errors is much more reliable than using only the WER deletion errors because not all deletion errors are produced by missing words; a number of WER deletions appears due to reordering errors. Information about the base form is used in order to eliminate inflectional errors. For the example in Table 6.1, the word "can" will be identified as missing.

Extra words

Analogously to missing words, extra words are also detected from the WER and PER errors; the words considered as extra are those which occur as insertions in WER errors and at the same time occur only as hypothesis PER errors without sharing the base form with any reference error.

$$N(extra) = \sum_{k=1}^{K} \sum_{eb \in hberr_k} n(e, inserr_k)$$

where $inserr_k$ is the set of insertion WER errors and $hberr_k$ is the set of base form HPER errors. In the example in Table 6.1, none of the words will be classified as an extra word.

Incorrect lexical choice

The words in the reference translation which are classified neither as inflectional errors nor as missing words are considered as incorrect lexical choices:

$$N(lex) = \sum_{k=1}^{K} \sum_{eb} n(eb, rberr_k) - N(miss)$$

As in the case of the inflectional errors, a definition using hypothesis errors and extra words is also possible, but in this work we choose to use reference errors.

In Table 6.1 the word “Mister” in the reference (or the word “Mrs” in the hypothesis) is considered to be translated by an incorrect lexical choice.

6.3 Experimental results

In order to compare the results of the proposed automatic error analysis with human evaluation, the methods described in the previous sections are applied to several translation outputs with the available results of human error analysis. The translation outputs were produced in the framework of the GALE project and the TC-STAR project.

For comparing different translation systems, new error rates based on the error categories are introduced and compared with standard error rates WER, PER and TER. These error rates are then calculated for various translation outputs generated by the translation systems described in Sections 4.3 and 5.3 in order to investigate how the new metrics reflect the effects of data sparseness and morpho-syntactic transformations. An additional experiment on five outputs generated in the second TC-STAR evaluation by five distinct translation systems is performed.

6.3.1 Comparison with the results of human error analysis

The GALE corpora considered in this analysis consist of Arabic-to-English broadcast news (BN) translation and Arabic-to-English and Chinese-to-English newswire (NW) translations. The TC-STAR corpora consist of three Spanish-to-English and three English-to-Spanish translated transcripts of European Parliament plenary sessions (EPPS), two Final Text Editions (FTE) and one Verbatim Transcription (VT). The translation of all the texts was performed using state-of-the-art statistical phrase-based machine translation systems. It should be noted that for all TC-STAR data the same training corpus consisting of FTE texts is used, which produces a slight mismatch for the translation of VT data.

The results of both human and automatic error analysis for all analysed texts are presented in the following sections. In addition, the Pearson (r) and Spearman rank (ρ) correlation coefficients between human and automatic results are calculated. Both coefficients assess how well a monotonic function describes the relationship between two variables, but the Spearman correlation does not require a linear relationship between the variables. The Spearman's rank correlation coefficient is equivalent to the Pearson correlation on ranks. A Pearson correlation of +1 means that there is a perfect positive linear relationship between variables, and a Spearman correlation of +1 that the ranking using both variables is exactly in the same order. A correlation of -1 means that there is a perfect negative relationship between variables i.e. exactly inverse ranking. A correlation of 0 means there is no relationship between the two variables. Thus, the higher value of r and ρ, the more similar are the metrics.

Human error analysis: Human error analysis and classification is a time-consuming and difficult task, and it can be done in various ways. For example, in order to find errors in a translation output it can be useful to have one or more reference translations. However, there are often several correct translations of a given source sentence and some of them might not correspond to the reference translations, which poses difficulties for evaluation and error analysis. The errors can be counted as an exact comparison between references and translation outputs which is then very similar to the automatic error analysis. However, much more flexibility can be allowed and use references only for the semantic aspect, i.e. allow substitution of words and expressions by synonyms, syntactically correct word order, etc. There are also other aspects which may differ between human evaluations, for example counting each problematic word as an error or counting groups of words as one error, etc. Furthermore, human error classification is definitely not unambigous; often it is not easy to determine exactly in which particular error category some error belongs, and variations between different human evaluators are possible. For the error categories described in the previous sections, especially difficult is disambiguating between incorrect lexical choice and missing words or extra words. For example, if the translation output is "the day before yesterday" and translation reference is "yesterday", it could be considered as an incorrect lexical choice, but also as a group of extra words. Similarly, there are several interpretations of errors if "the one who will come" is translated as "which comes".

In this work, for the GALE corpora the errors are classified by two human annotators with respect to a given reference. This kind of error analysis is basically carried out in a similar manner as the automatic error analysis. For the TC-STAR corpora, error analysis is performed by three human annotators taking the reference translations into account only from the semantic point of view [Vilar & Xu$^+$ 06]. This type of error analysis is much less strict, i.e. it identifies many fewer words as errors, as can be seen in Table 6.6.

Table 6.6: Examples of two variants of human error analysis, with and without respect to a given reference translation; the marked errors are detected with respect to the reference, whereas no errors are detected if the reference translation is considered only for the semantics.

reference translation	obtained output
we celebrated the fifteenth anniversary	we **have held** the fifteenth anniversary
I think this is a good moment	I **believe that** this is a good **opportunity**
to achieve these ends	**for** these **purposes**
in 2002	in **the year** 2002
in Europe we must also learn	**also** in Europe we must learn

Results on the GALE corpora

For the GALE corpora, translation errors are classified both by humans and by automatic tools in one of the following categories: inflectional errors, reordering errors, missing words, extra words and incorrect lexical choice. In addition, distribution over main POS classes — nouns (N), verbs (V), adjectives (A), adverbs (ADV), pronouns (PRON), determiners (DET), prepositions (PREP), conjunctions (CON), numerals (NUM) and punctuation marks (PUN) — is estimated.

The results of both the human and the automatic error classification are shown in Table 6.7. The number of errors in each category is normalised over the total number of errors and the percentage is presented. Both results show the same tendencies: that for the Arabic-to-English Broadcast News translation the main sources of errors are extra words and incorrect lexical choice; for the Newswire corpus the predominant problem is incorrect lexical choice; and for the Chinese-to-English the majority of errors are caused by missing words, incorrect lexical choice and wrong word order.

Table 6.7: Results of human and automatic error analysis for the GALE corpora.

	human					automatic				
output	infl	order	miss	ext	lex	infl	order	miss	ext	lex
Ar-En BN	5.0	9.6	19.8	31.8	33.8	5.3	15.1	14.5	31.4	33.7
Ar-En NW	6.1	8.3	26.8	20.2	38.6	6.4	10.9	27.4	20.3	35.0
Cn-En NW	5.1	16.9	38.3	12.6	27.1	4.9	21.1	29.9	14.4	29.7

The results for the ten basic POS classes are shown in Table 6.8. Again, from both human and automatic error analysis the same conclusions can be drawn; the verbs are the main source of inflectional errors for Arabic–English translation whereas for the Chinese–English translation the majority of inflectional errors is produced by nouns. As for the missing words, for Arabic–English translation the verbs are again the most problematic category, followed by nouns, pronouns and prepositions. For the Chinese–English translation the majority of missing words are nouns, then verbs and prepositions. For both NW corpora a large number of extra punctuation is present. Prepositions are problematic as extra words in all cases, as well as de-

Table 6.8: Results of human and automatic error analysis for the GALE corpora: distribution of different error types over basic POS classes.

(a) Arabic–English Broadcast News

Ar-En BN		V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	hum	75.0	15.0	0	0	10.0	0	0	0	0	0
	aut	74.0	13.0	0	0	13.0	0	0	0	0	0
miss	hum	36.7	12.7	5.0	7.6	13.9	3.8	10.1	1.3	1.3	7.6
	aut	23.8	22.2	3.2	6.4	14.3	3.2	7.9	1.6	0	17.4
extra	hum	8.7	15.0	3.2	7.1	6.3	26.0	19.7	5.3	2.4	6.3
	aut	8.0	16.8	2.2	5.8	10.9	26.3	15.3	3.1	3.6	8.0
lex	hum	16.3	17.0	5.2	5.2	12.6	3.7	17.8	7.4	4.4	10.4
	aut	21.8	15.0	4.8	6.1	10.9	3.4	16.3	4.8	4.1	12.8

(b) Arabic–English Newswire

Ar-En NW		V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	hum	81.8	9.2	4.5	0	4.5	0	0	0	0	0
	aut	75.0	8.4	4.2	0	12.4	0	0	0	0	0
miss	hum	25.8	14.4	4.1	5.2	18.5	9.3	15.4	5.2	0	2.1
	aut	34.3	11.8	2.9	4.0	13.7	9.8	13.7	3.0	0	6.8
extra	hum	12.5	16.4	2.7	1.4	9.6	13.7	16.4	6.8	0	20.5
	aut	14.5	22.4	2.6	3.9	5.3	10.5	13.2	6.6	0	21.0
lex	hum	27.1	17.1	2.9	5.7	15.7	5.7	16.4	5.0	1.4	3.0
	aut	20.6	16.8	5.3	6.9	17.6	5.3	13.1	7.6	1.5	5.3

(c) Chinese–English Newswire

Cn-En NW		V	N	A	ADV	PRON	DET	PREP	CON	NUM	PUN
infl	hum	36.8	63.2	0	0	0	0	0	0	0	0
	aut	40.0	60.0	0	0	0	0	0	0	0	0
miss	hum	17.0	26.0	4.2	4.9	5.9	8.3	17.4	8.0	1.7	6.6
	aut	18.6	30.2	2.9	5.4	5.4	6.2	14.9	5.4	2.0	9.0
extra	hum	6.3	18.9	5.4	1.0	2.1	22.1	24.2	5.3	0	14.7
	aut	5.1	32.5	7.7	0	0.9	17.1	20.5	1.7	3.4	11.1
lex	hum	10.3	42.9	6.4	3.4	2.5	6.4	15.8	4.4	2.0	5.9
	aut	19.5	29.9	5.4	4.6	4.6	4.2	15.2	6.2	2.5	7.9

terminers and nouns. For all corpora, the majority of incorrect lexical choices belongs to nouns, verbs and prepositions, and for the Arabic-to-English translation pronouns as well.

Table 6.9 presents correlations between the results of the automatic and the human analysis. The correlation function is defined by percentage of errors in each category (first column) or by the percentage of errors for each POS class within a particular error category. It can be seen that the automatic measures have very high correlation coefficients with respect to the results of human evaluation. Correlations for the inflectional error category are higher than for the other categories, which can be explained by the fact mentioned in the previous sections

Table 6.9: Correlation coefficients for the GALE corpora: Spearman rank ρ (left column) and Pearson r (right column) coefficient.

output	error categories		distribution of errors over POS classes							
			infl		miss		extra		lex	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
Ar-En BN	0.900	0.955	0.997	0.999	0.918	0.790	0.870	0.947	0.924	0.924
Ar-En NW	1.000	0.994	0.979	0.994	0.921	0.922	0.912	0.916	0.894	0.960
Cn-En NW	1.000	0.930	1.000	0.998	0.927	0.973	0.879	0.853	0.788	0.914

that disambiguation between missing words, extra words and incorrect lexical choice is often difficult, both for humans and for machines.

Results on the TC-STAR corpora

The experiments on the TC-STAR corpora are similar to those on the GALE corpora. However, there are some differences since human error classification is carried out in a somewhat different way. The error categories considered are inflectional errors, missing words, reordering errors and incorrect lexical choice, i.e. the same as for the GALE experiments except extra words. The distribution of errors over POS tags is not analysed on these corpora, but the following details about inflectional errors are investigated: verb tense errors, verb person errors, adjective gender errors and adjective number errors. Correlation coefficients are calculated both for general error categories and for inflectional details.

The results of the error classification are shown in Table 6.10. Human and automatic error analysis again produce similar results; the majority of errors are caused by incorrect lexical choice, whereas for the Spanish output the amount of inflectional errors is also very high due to the richer morphology of the Spanish language. The number of reordering errors is higher in English outputs which can be explained by the more flexible rules for word order in the Spanish language. The percentage of missing verbs is significantly higher than for Spanish. The probable reason for this is the nature of Spanish verbs. Since person and tense are contained in the suffix, Spanish pronouns are often omitted, and auxiliary verbs do not exist for all tenses. This could be problematic for a translation system, because it processes only one Spanish word which actually contains two (or more) English words.

Table 6.11 presents results for inflectional details about verbs and adjectives, i.e. tense, person, gender and number. Both human and automatic error analysis indicate that the most problematic inflectional category is the tense of verbs, especially for translation into Spanish having an almost two times higher error rate than for English. This is due to the very rich morphology of Spanish verbs; one base form might have up to about forty different inflections.

Correlation coefficients are shown in Table 6.12. It can be seen that for this corpus, correlations for error categories are lower than for the GALE corpus although all are rather high, above 0.5. This is due to the flexible human evaluation which is carried out on this corpora, i.e. without taking the reference translation strictly into account. However, for the inflectional error analysis the correlations are very high, above 0.9.

Table 6.10: Results of human and automatic error analysis for the TC-STAR corpora.

output	human				automatic			
	infl	order	miss	lex	infl	order	miss	lex
Es-En1 FTE	10.5	21.5	27.3	40.7	10.9	22.1	19.5	47.5
Es-En2 FTE	18.2	18.9	31.1	31.8	13.8	21.9	15.7	48.6
Es-En1 VT	12.7	22.8	21.2	43.3	14.2	18.4	20.0	47.4
En-Es1 FTE	31.7	16.0	20.6	31.7	23.7	17.4	18.5	40.4
En-Es2 FTE	34.8	15.9	18.4	30.9	22.1	18.9	12.4	46.6
En-Es1 VT	30.5	11.5	26.8	31.2	20.8	16.6	17.4	45.2

Table 6.11: Results of human and automatic error analysis for the TC-STAR corpora – inflectional details.

output	human				automatic			
	Vten	Vper	Agen	Anum	Vten	Vper	Agen	Anum
Es-En1 FTE	8.1	2.4	0	0	7.9	3.0	0	0
Es-En2 FTE	14.9	3.3	0	0	9.9	3.9	0	0
Es-En1 VT	8.5	4.2	0	0	10.4	3.8	0	0
En-Es1 FTE	15.6	8.5	4.3	3.3	11.7	7.4	2.4	2.2
En-Es2 FTE	15.0	8.7	5.8	5.3	10.5	5.8	3.0	2.8
En-Es1 VT	13.4	8.6	4.8	3.7	11.2	7.0	1.3	1.3

Table 6.12: Correlation coefficients for the TC-STAR corpora: Spearman rank ρ (left column) and Pearson r (right column).

output	error categories		infl. errors over POS classes	
	ρ	r	ρ	r
Es-En1 FTE	0.800	0.935	1.000	0.996
Es-En2 FTE	0.800	0.552	1.000	0.983
Es-En1 VT	0.800	0.978	1.000	0.991
En-Es1 FTE	0.950	0.754	1.000	0.987
En-Es2 FTE	0.600	0.572	1.000	0.998
En-Es1 VT	1.000	0.538	0.950	0.990

6.3.2 Comparison of translation systems

In order to compare translation outputs generated by different translation systems using the proposed error categories, we introduce word error rates for each error category, namely we normalise the number of errors over the reference length. These error rates are defined in Section B.3. For each of the translation outputs, all novel evaluation metrics are calculated and

then compared. Three main aspects are considered: languages, size of the training corpus and the use of morpho-syntactic transformations.

An overview about how these metrics behave in comparison with the standard word error rates WER, PER and TER is presented in Table 6.13. All error rates are calculated on the analysed translation outputs described in Section 6.3.1. It can be seen that the sum of all error categories ΣER is always greater than PER, lower than WER and similar to (although in the majority of cases lower than) TER.

Table 6.13: Number of errors in five error categories normalised over reference length and compared with standard word error rates WER, PER and TER.

%	WER	PER	TER	INFER	RER	MSER	EXER	LXER	ΣER
Ar-En (BN)	29.6	22.4	28.0	1.46	4.18	3.99	8.68	9.31	27.6
Ar-En (NW)	18.8	14.5	17.8	1.14	1.95	4.86	3.62	6.24	17.8
Cn-En (NW)	34.6	21.4	30.0	1.58	6.77	9.58	4.63	9.53	32.1
Es-En1 (FTE)	34.5	24.7	32.1	1.63	6.15	5.44	2.13	13.2	28.6
Es-En2 (FTE)	37.6	26.3	34.9	2.91	6.36	4.55	3.84	14.1	31.8
Es-En (VT)	42.2	30.8	39.8	3.05	6.23	6.77	3.47	16.0	35.5
En-Es1 (FTE)	42.8	31.7	39.7	5.29	6.84	7.27	2.54	15.9	37.8
En-Es2 (FTE)	40.4	29.8	37.9	4.20	6.45	4.24	4.51	15.9	35.3
En-Es (VT)	46.5	35.0	44.1	4.80	7.09	7.45	2.71	19.3	41.3

Spanish–English language pair

Table 6.14 presents the results for the following training set-ups:

- training on the full bilingual corpus – a large task-specific corpus (1.3M);
- training on a small task-specific corpus (13k);
- training on a very small task-specific corpus (1k);
- training only on a conventional dictionary.

The effects of the local reorderings are investigated for each size of the training corpus. For the dictionary, the effects of reducing adjectives and OOV words into the base forms are also examined.

Several things can be observed. As can be expected, the errors caused by incorrect lexical choice grow most significantly when reducing the training corpus. The decrease in the training corpus size also increases the number of reordering errors. This can be explained by the fact already discussed in the previous secions, i.e. that the phrase-based translation system is able to generate frequent noun-adjective groups in the correct word order, but unseen or rarely seen groups introduce difficulties. This is also the reason why the reduction of reordering errors by applying local POS-based reorderings is more significant for the small training corpora. Furthermore, it can be seen that it is hard to find any correspondence between inflectional errors in the English output and the size of the training corpus. A probable reason is that the morphology

Table 6.14: Comparison of different versions of the TC-STAR system: different sizes of the training corpus with and without local POS-based reorderings.

(a) Spanish–English

%	INFER	RER	MSER	EXER	LXER	ΣER
1.3M	1.82	5.54	4.13	3.55	14.4	29.5
+reorder adjectives	1.74	5.13	4.02	3.87	14.8	29.5
13k	2.05	7.36	4.52	3.86	18.8	36.6
+reorder adjectives	1.75	5.68	4.39	4.08	18.5	34.4
1k	1.65	7.86	5.04	4.78	28.1	47.4
+dictionary	1.99	8.14	5.98	3.35	23.6	43.1
+reorder adjectives	1.91	6.08	5.25	4.33	23.3	40.8
dictionary	1.75	7.79	5.90	4.06	35.9	55.4
+reorder adjectives	1.75	6.12	6.04	4.39	35.5	53.8
+adjective base	1.79	6.44	5.97	4.35	33.6	52.1
+OOV base	3.91	6.86	5.85	4.53	28.6	49.7

(b) English–Spanish

%	INFER	RER	MSER	EXER	LXER	ΣER
1.3M	4.81	6.06	5.31	3.71	15.7	35.5
+reorder adjectives	4.70	5.95	5.23	3.61	15.5	35.0
13k	5.91	8.23	5.53	3.97	20.2	43.9
+reorder adjectives	5.54	7.47	5.62	3.82	20.0	42.5
1k	6.50	8.78	5.79	4.42	28.7	54.2
+dictionary	6.86	9.03	5.56	4.27	24.8	50.5
+reorder adjectives	7.28	8.14	6.10	4.05	24.7	50.2
dictionary	8.54	8.49	6.10	3.68	33.8	60.6
+reorder adjectives	8.15	7.72	6.45	3.88	33.3	59.5

is not particularly rich in the English language so that the reduction of the training corpus introduces many more incorrect lexical choices than inflectional errors. On the other hand, for the Spanish corpus an increase in inflectional errors can be seen, which is particularly high when the dictionary is the only training material. It can also be noted that introducing a dictionary as an additional tranining corpus decreases the number of incorrect lexical choices but increases the number of inflectional and reordering errors. As for missing words and extra words, it is hard to find any relation either to the size of the training corpus or to the local reorderings. As for morphological transformations for translation with the dictionary, reducing adjectives to base forms decreases the number of incorrect lexical choices, and reducing all OOV words to the base forms further improves this error rate significantly. However, this transformation increases the number of inflectional errors, but this loss is much smaller than the gain with the incorrect lexical choice so the overall performance is improved.

For the full training corpus, we investigate more details about POS-based reorderings on two subsets of the test corpus: reordered sentences and the rest, as described in Section 4.3. For this analysis, we use the overall RER measure as well as RER of noun-adjective groups and RER

Table 6.15: Effects of local POS-based reorderings for Spanish–English translation: error rates for reordered sentences and for the rest.

(a) English output

Spanish→English		RER	RER (N,A)	RER (V)
reordered	baseline	5.99	2.64	0.87
	reorder adjectives	5.43	2.07	0.91
not reordered	baseline	4.19	1.26	1.02
	reorder adjectives	4.27	1.27	1.05

(b) Spanish output

English→Spanish		RER	RER (N,A)	RER (V)
reordered	baseline	6.49	2.34	0.51
	reorder adjectives	6.33	2.24	0.54
not reordered	baseline	4.61	1.15	0.63
	reorder adjectives	4.66	1.23	0.60

of verbs. The results in Table 6.15 show that the overall RER of the reordered set is decreased by the local reorderings whereas for the rest of the sentences a small increase can be observed. Furthermore, it can be noted that for the reordered set the RER of verbs is significantly smaller than the RER of nouns and adjectives which has been improved by local reorderings. For the rest of the sentences there are no significant differences either between RERs of different POS groups or between the system with reorderings and the baseline system. The same tendencies occur for the other translation direction.

German–English language pair

For this language pair, the following set-ups are analysed for the EUROPARL corpus:

- training on the full bilingual corpus – a large task-specific corpus (700k);
- training on a very small task-specific corpus (1k);
- training only on a conventional dictionary.

For each configuration, the effects of long-range verb reorderings and compound spliting are investigated.

Table 6.16 shows that as in the case of Spanish–English translation, incorrect lexical choice is the category which depends most on the size of the training corpus. Unlike for the Spanish–English pair, the reordering error rate does not change when decreasing the corpus size. This is an explanation for the experiments reported in Section 5.3 where we stated that long-range reorderings do not have more impact on the small corpora than on the large ones. As can be seen, RER is improved by the reorderings more or less equally for all corpora. Splitting compounds on the other hand is more useful for the small corpora and reduces the number of lexical errors.

Table 6.16: Comparison of different versions of the EUROPARL system: different sizes of the training corpus with and without long-range POS-based reorderings and compound word splitting.

%	INFER	RER	MSER	EXER	LXER	ΣER
700k	2.88	14.7	8.71	7.88	29.8	63.9
+reorder verbs	2.88	13.7	8.25	7.99	29.8	62.6
+split compounds	2.91	13.8	8.04	7.89	29.7	62.4
1k	2.45	13.0	8.00	8.30	46.1	77.8
+dictionary	3.13	14.2	8.46	7.87	40.4	74.1
+reorder verbs	3.13	13.9	8.33	7.77	40.2	73.3
+split compounds	3.20	14.0	8.10	7.88	39.3	72.6
dictionary	3.02	13.8	8.42	7.81	44.1	77.2
+reorder verbs	3.11	13.3	8.65	7.66	43.2	75.9
+split compounds	3.14	13.6	8.27	7.98	42.6	75.6

Table 6.17: Effects of long-range POS-based reorderings for German–English translation: error rates for reordered sentences and for the rest.

(a) English output

German→English		RER	RER (N,A)	RER (V)
reordered	baseline	15.6	4.45	3.16
	reorder verbs	14.4	4.29	2.63
not reordered	baseline	8.58	3.77	0.65
	reorder verbs	8.68	3.74	0.67

(b) German output

English→German		RER	RER (N,A)	RER (V)
reordered	baseline	14.4	3.84	1.82
	reorder verbs	13.8	3.76	1.61
not reordered	baseline	10.7	3.35	1.25
	reorder verbs	10.5	3.17	1.13

Similar to the Spanish–English pair, more details about reorderings are examined for the large training corpus, and the overall RER along with the RER of noun-adjective groups and verbs are reported in Table 6.17. The overall RER of the reordered set is much higher than for the rest, and is significantly improved by reorderings. Reordering errors of noun-adjective groups and verbs have a similar value, in contrast to the Spanish–English language pair where the RER of nouns and adjectives is much higher than for verbs. The long-range verb reorderings lead to a decrease in the verb RER, and also to a small derease in the noun-adjective RER. As already mentioned in Section 4.3, long-range reorderings introduce both direct and indirect improvements of the system. For the rest of the sentences, small improvements of all reordering error rates can be observed. For the other translation direction similar phenomena can be perceived, although all improvements are smaller than for translation into English.

Serbian–English language pair

For the Serbian–English translation, the following systems are analysed:

- training on the full (small) bilingual corpus – 2.6k;
- training on the extremely small bilingual corpus – 0.2k.

The effects of reducing words into the base forms and the verb treatment are also explored for both systems.

As in the case of the other two language pairs, corpus size has the largest influence on incorrect lexical choices (Table 6.18). However, a significant increase in missing words can be observed for the extremely small training corpus. Reducing all words into base forms leads to large improvements in lexical error rate LXER, but at the same time to an increase in inflectional errors INFER: most Serbian inflections are not relevant for the translation into English, but some are. Gender and case of nouns and adjectives are completely redundant, but number of nouns is important for distinguishing between singular and plural. The most important inflections are person and tense of verbs; as explained in Section 5.2, they are expressed via suffix and the pronoun is often omitted. Further analysis of inflectional errors for over POS classes showed that the verbs are indeed the main source of this increase; for the baseline system INFER of verbs is 1.68%, and for the system with base forms it reaches 3.63%. The verb treatment overcomes this problem to some extent, and for the full corpus further reduces the number of incorrect lexical choices. For the extremely sparse corpus, however, there are no changes in LXER, but a large reduction in the number of missing words can be observed. In order to better understand this phenomenon, a further analysis of missing words over POS classes is carried out and the results are presented in Table 6.19. It can be seen that the verb treatment significantly reduces the number of missing verbs and missing pronouns.

Table 6.18: Comparison of different versions of Serbian–English translation system: different sizes of the training corpus with and without morpho-syntactic transformations.

%	INFER	RER	MSER	EXER	LXER	ΣER
2.6k	2.59	5.01	8.68	3.19	23.9	43.4
+base forms	5.70	4.19	7.82	3.93	20.1	41.7
+verb treatment	5.57	5.36	7.94	3.54	18.1	40.5
0.2k	1.60	4.06	12.1	2.07	45.1	65.0
+base forms	4.54	3.24	12.5	2.63	38.1	61.0
+verb treatment	3.97	5.14	7.90	5.70	38.7	61.7

Comparison of different translation outputs generated in the TC-STAR evaluation

For all translation outputs analysed in the previous sections, the same phrase-based translation system is used for all experiments. In order to examine how the new error rates reflect the differences between distinct translation systems, we carried out an error analysis of the different translation outputs generated by five distinct translation systems in the second TC-STAR evaluation. A total of nine different systems participated in the evaluation, and we selected

Table 6.19: Analysis of missing words for Serbian–English translation system trained on an extremely small corpus with and without morpho-syntactic transformation.

		0.2k	+bases	+verbs
MSER	V	4.2	4.2	**2.5**
	N	1.2	1.2	0.9
	A	0.1	0.2	0.1
	ADV	0.9	1.0	0.8
	PRON	2.5	2.5	**1.0**
	DET	2.5	2.5	2.1
	PREP	0.6	0.7	0.5
	CON	0.1	0.2	0.1
	NUM	0	0	0
	PUN	0	0	0

five representative systems for our experiments which will be referred to as A, B, C, D and E. For the English language we used the outputs of four systems A, B, C, and D, and for Spanish additionaly the output of a system E.

In Table 6.20 the new error rates for all translation outputs are presented along with the BLEU score as the official metric of the evaluation. For translation into English, systems A, B and C have very similar BLEU scores as well as all error categories. The worst-ranked system according to the BLEU is system D, and from the error rates it can be seen that the main problem for this system is incorrect lexical choice. The number of extra words is also larger for this system than for the others, and the number of reordering errors too.

Table 6.20: Error categories for different translation systems.

(a) English outputs

English	BLEU	INFER	RER	MSER	EXER	LXER	ΣER
A	53.5	2.47	5.82	4.72	4.07	14.5	31.6
B	53.1	2.30	5.68	4.84	3.50	13.9	30.2
C	52.8	2.10	5.93	5.58	3.20	14.1	30.9
D	45.4	2.64	6.87	3.69	5.18	17.5	35.9

(b) Spanish outputs

Spanish	BLEU	INFER	RER	MSER	EXER	LXER	ΣER
A	50.0	4.78	5.62	4.65	4.12	15.2	34.4
B	48.2	4.80	5.80	5.31	3.78	15.1	34.8
C	49.6	4.93	5.62	5.33	3.11	14.7	33.7
D	38.9	5.48	6.72	4.70	4.39	19.4	40.7
E	38.6	5.11	7.75	4.62	4.69	19.0	41.2

For translation into Spanish, the BLEU scores are similar for the three systems A, B and C and for the two systems D and E they are lower. The error rates show that the main differences

between systems A, B, C on the one hand and systems D and E on the other are incorrect lexical choices. The number of inflectional and reordering errors is also higher for the system D and E.

Since the largest difference between systems for both translation directions is observed for lexical error rate LXER, a further analysis of this error category is carried out, namely distribution of errors over POS classes, and the results are shown in Table 6.21. For the English output, the main differences between system D and the others are incorrect lexical choices of prepositions and nouns, although a notable difference can be observed also for the other POS classes. Similar tendencies can be seen for the Spanish output, and in addition the difference for the verbs seems to be more significant.

Table 6.21: Incorrect lexical choice of different POS classes produced by different translation systems.

(a) English outputs

English		A	B	C	D
LXER	V	3.98	4.00	3.81	4.26
	N	3.38	3.18	3.31	**4.20**
	A	1.00	1.03	1.00	1.24
	ADV	0.92	0.94	1.01	1.26
	PRON	1.50	1.46	1.42	1.91
	DET	0.82	0.69	0.85	0.99
	PREP	1.97	1.90	1.87	**2.44**
	CON	0.28	0.24	0.27	0.40
	NUM	0.10	0.11	0.11	0.11
	PUN	0.50	0.37	0.41	0.49

(b) Spanish outputs

Spanish		A	B	C	D	E
LXER	V	3.71	3.63	3.53	**4.11**	**4.27**
	N	3.25	3.22	3.16	**5.08**	**4.44**
	A	1.29	1.34	1.25	1.72	1.83
	ADV	0.84	0.84	0.81	1.10	1.08
	PRON	0.71	0.79	0.76	0.99	0.84
	DET	1.39	1.37	1.38	1.44	1.62
	PREP	2.92	2.94	2.89	**3.77**	**3.67**
	CON	0.50	0.46	0.51	0.62	0.57
	NUM	0.08	0.11	0.06	0.07	0.09
	PUN	0.44	0.33	0.34	0.48	0.53

For Spanish outputs, a detailed analysis of inflectional errors is performed too, because Spanish morphology is more problematic than English, and the results are shown in Table 6.22. The worst-ranked systems D and E produce more inflectional errors for nouns, pronouns and especially determiners. For adjectives and adverbs there are no significant differences between these systems and the other three systems, and for verbs system E shows the best performance.

This shows that although overall performance of some system is worse than for the others, this system still can outperform the others in some particular aspects.

Table 6.22: Inflectional errors for different POS classes in Spanish outputs produced by different translation systems.

Spanish		A	B	C	D	E
INFER	V	2.15	2.13	2.23	2.14	**1.95**
	N	0.28	0.28	0.30	0.46	0.44
	A	0.54	0.50	0.56	0.54	0.54
	ADV	0.01	0.01	0.02	0.02	0.03
	PRON	0.29	0.19	0.28	0.39	0.31
	DET	1.50	1.67	1.48	**2.02**	**1.81**

6.4 An alternative approach to automatic error analysis

This section presents another approach to automatic error analysis of translation output also based on the standard word error rates and linguistic knowledge. The main idea of this method is to calculate WER and PER separately for each word class, and then to perform the further analysis. In the next sections a detailed description of this approach will be presented along with some experimental results. The general scheme is presented in Figure 6.2.

6.4.1 Word error rates of each POS class

Another way to estimate POS-based error rates is to create a new reference and a new hypothesis for each POS class by extracting all words belonging to this class, and then to calculate the standard WER and PER. The obtained error rates are then weighted with the relative frequency of the respective class. These error rates will be referred to as WER′ and PER′.

From the example in Table 6.1 (page 56), six references and six hypotheses will be created: for nouns, verbs, adverbs, pronouns, numerals and punctuation marks. The new references and hypotheses are shown in Table 6.23. The WER′ and PER′ of adverbs, pronouns, numerals and punctuations are equal to zero. For nouns, the standard WER and PER are 25% and for verbs are 100%. After weighting with the relative frequencies of the corresponding POS classes, the final error rates are $\text{WER}'(\text{N}) = \text{PER}'(\text{N}) = (4/12) \cdot 25\% = 8.3\%$ and $\text{WER}'(\text{V}) = \text{PER}'(\text{V}) = (2/12) \cdot 100\% = 16.7\%$. It can be noted that the sum of PER' over all POS classes is equal to the standard PER whereas the sum of WER' is less than standard WER. This is because part of the information about word order is lost by creating separate references and hypotheses.

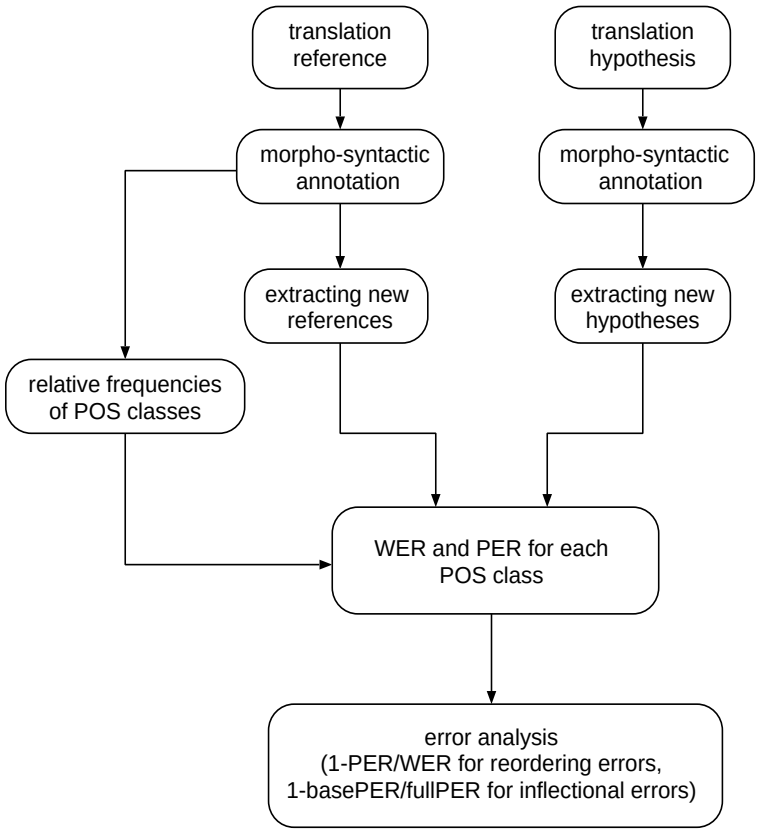

Figure 6.2: Error analysis based on standard word error rates calculated separately for each word class.

Table 6.23: Example of new references and hypotheses for each POS class.

POS class	reference	hypothesis
noun	Mister#N Commissioner#N hours#N time#N	Mrs#N Commissioner#N hours#N time#N
verb	can#V be#V	is#V
adverb	sometimes#ADV too#ADV	sometimes#ADV too#ADV
pronoun	much#PRON	much#PRON
numeral	twenty-four#NUM	twenty-four#NUM
punctuation	,#PUN .#PUN	,#PUN .#PUN

6.4.2 Inflectional errors

Estimation of inflectional errors is based on the PER of full forms and the PER of base forms: the relative difference between these PERs is calculated. The larger this difference is, more inflectional errors are present. This difference can be calculated for all words as well as for different POS classes as described above.

The overall amount of inflectional errors for the example presented in Table 6.1 is $1 - \frac{2/12}{3/12} = 33.3\%$. From Table 6.23, the relative difference for nouns is equal to zero because

the PER of the full forms is equal to the PER of the base forms. For verbs, the relative difference is 50%: the full form PER is 100% and the base form PER is 50% because the words "is" and "be" have the same base form "be".

This method does not give an exact number of inflectional errors like the method presented in Section 6.2, but gives a general overview of the amount of inflectional errors in general, or for particular POS classes.

6.4.3 Reordering errors

The overall amount of reordering errors is estimated using the relative difference between WER and PER: the larger this difference is, more reordering errors are present. For the example in Table 6.1, the overall relative difference is $1 - 25/41.7 = 40.0\%$. As for the case of inflectional errors, reordering errors can be also estimated for particular POS classes using the relative difference between WER$'$ and PER$'$.

Similar to the methods for inflectional errors, the method described in Section 6.2 gives the number of reordering errors in the translation output, whereas this method gives a general overview of the amount of reordering errors.

6.4.4 Experimental results

Inflectional and reordering errors are estimated on the Spanish, English and German translation outputs described in Section 4.3. Table 6.24 presents the results for the Spanish and English TC-STAR outputs and Table 6.25 for the German and English outputs from the EUROPARL corpus. $\Delta(\text{WER},\text{PER})$ denotes the relative difference between overall WER and PER, whereas $\Delta(\text{WER}'(N,A),\text{PER}'(N,A))$ and $\Delta(\text{WER}'(V),\text{PER}'(V))$ are relative WER$'$ $-$ PER$'$ differences for noun-adjective groups and verbs respectively. The inflection errors are represented by the relative difference between overall standard PER of full forms and of base forms PER(b).

Table 6.24: Relative differences for translation outputs generated by Spanish–English translation systems with and without local adjective reorderings.

output	English		Spanish	
system	baseline	reorder	baseline	reorder
$\Delta(\text{WER},\text{PER})$	28.7	27.8	24.4	23.5
$\Delta(\text{WER}'(N,A),\text{PER}'(N,A))$	26.6	24.7	24.4	22.4
$\Delta(\text{WER}'(V),\text{PER}'(V))$	8.9	8.4	3.1	3.0
$\Delta(\text{PER},\text{PER}(b))$	8.8	8.8	15.6	15.7

The following observations can be noted: inflectional errors are much more present in the Spanish and in the German output than in the English ones. These errors are not influenced by POS-based word reorderings. The overall amount of reordering errors is reduced by applying the corresponding reorderings for all outputs. For the Spanish–English language pair, the

Table 6.25: Relative differences for translation outputs generated by German–English translation systems with and without long-range verb reorderings.

output	English		German	
system	baseline	reorder	baseline	reorder
$\Delta(\text{WER}, \text{PER})$	30.9	29.7	25.3	24.7
$\Delta(\text{WER}'(N,A), \text{PER}'(N,A))$	12.3	11.8	14.7	14.3
$\Delta(\text{WER}'(V), \text{PER}'(V))$	8.7	3.8	6.0	5.7
$\Delta(\text{PER}, \text{PER}(b))$	5.6	5.6	11.2	11.1

number of reordering errors involving nouns and adjectives is much higher than those involving verbs. These errors are reduced by local reorderings whereas the verbs are not affected. For the German–English pair, long-range reorderings mainly reduce verb reordering errors, but a reduction can be seen for nouns and adjectives as well.

Further results concerning inflectional errors are presented in Figure 6.3: the distribution of inflectional errors over inflective POS classes for both English and Spanish output. The results obtained by the relative difference method are presented on the left. In order to compare these results with the INFER errors described in Section 6.2, the distribution of INFER over POS classes is shown on the right. A full line represents results for English and a dashed line for Spanish. Although the absolute numbers are different, the tendencies for both methods are the same: Spanish verbs, adjectives and determiners are causing the majority of inflectional errors, whereas for the English output the most problematic are the verbs and nouns. Apart from this, the amount of English verb inflectional errors is much lower than for Spanish, and for the nouns we see the opposite.

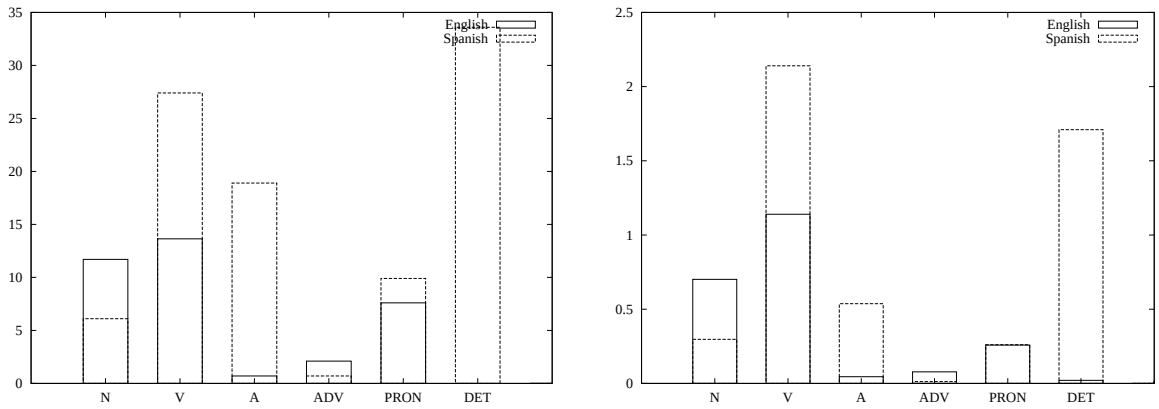

Figure 6.3: Inflectional errors [%]: relative difference $\Delta(\text{PER}', \text{PER}'(b))$ (left) and INFER (right) distributed over inflective POS classes for the English and Spanish TC-STAR outputs

The same results for the German and English EUROPARL outputs are shown in Figure 6.4. Again, the absolute numbers are different but the tendencies the same; for translation into German, the most problematic POS classes are adjectives and determiners, whereby a notable

amount of errors can be found in nouns, verbs and pronouns too. For the English output, the most problematic class is verbs.

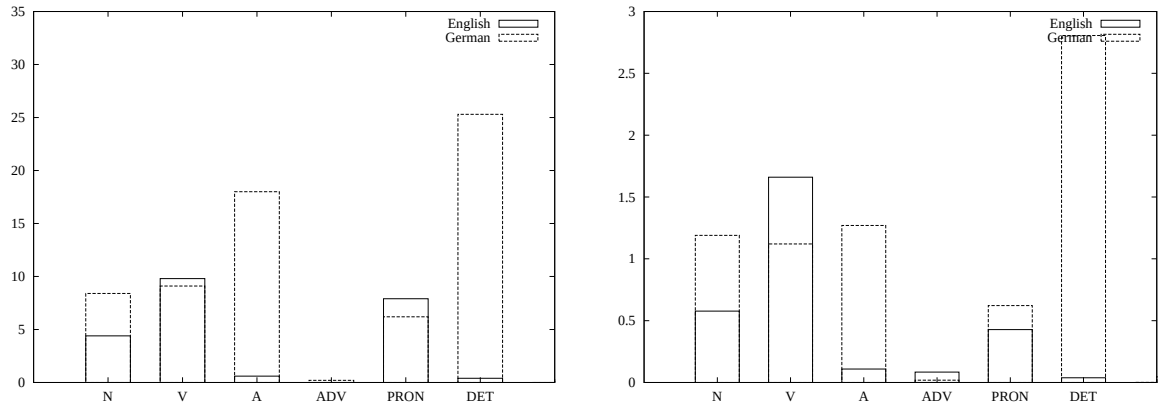

Figure 6.4: Inflectional errors [%]: relative difference $\Delta(\text{PER}', \text{PER}'(b))$ (left) and INFER (right) distributed over inflective POS classes for the English and German EUROPARL outputs.

6.5 Conclusions

A framework for automatic error analysis of machine translation output is proposed. The basic idea is to use details about actual errors extracted from the standard word error rates WER and PER in combination with linguistic knowledge in order to obtain more information about translation errors and to perform further analysis of particular phenomena. The overall goal is to get a better overview of the nature of actual translation errors – to obtain ideas about possible improvements of the translation system, to analyse the improvements achieved by particular methods, and to better understand the differences between distinct translation systems. There are many possibilities to carry out an automatic error analysis using the proposed methods. The focus of this work are five error categories: morphological (inflectional) errors, reordering errors, missing words, extra words and incorrect lexical choice. In addition, the distribution of these error types over POS classes is investigated. All new metrics can be applied to any language pair, the only prerequisite being the availability of a morpho-syntactic analyser for the target language.

The results of the proposed automatic methods are compared with the results obtained by human error analysis. Detailed experiments on different types of corpora and various language pairs are carried out and it is shown that the results of automatic error analysis correlate very well with the results of human analysis.

The new error measures can detect differences between different versions of the same phrase-based translation system. Some error categories are particularly sensitive to the amount of training data – mostly incorrect lexical choice, but also local reordering errors for the Spanish–English language pair and inflectional errors for highly inflected languages like Spanish. The improvements yielded by the POS-based reorderings are captured well by the new metrics

proposed for reordering errors. In addition, it is shown that the new measures are sensitive to the differences between distinct translation systems, i.e. that they can show what are the weak/strong points of particular systems.

The proposed metrics can be extended to other types of linguistic knowledge and other related phenomena, and also can be used for obtaining more particular details, for example examining the contributions of particular types of verb inflections, concordance of determiners and adjectives with nouns (e.g. for Spanish and German), errors of various named entities, etc.

One could think of using the TER measure instead of the basic edit distance measure WER; however, TER is an extension of WER including shift operations for handling reordering errors. Some of the metrics presented in this chapter also try to extend the WER measure in this direction (RER) so the advantages of TER could even become drawbacks for our task. This issue could be investigated in future work.

7 Scientific contributions

In this work, three aspects of the use of morpho-syntactic information for statistical machine translation have been systematically investigated on various tasks and different language pairs: POS-based local and long range word reorderings, translation with sparse bilingual training data and automatic error analysis of translation output.

POS-based word reorderings

Local and long-range word reorderings based on the source language POS tags have been introduced. The applied reordering rules are based on the knowledge about the sentence structure in the involved languages. Thus aligned bilingual corpora are not necessary. The transformations aim at the "harmonisation" of word orders in the two languages. In detail, the suggested reordering methods focus on the following aspects of structural differences: nouns and adjectives in the Spanish language, and verbs in the German language. Consistent improvements are achieved for both language pairs and all translation directions. A detailed evaluation showed that the reorderings especially improve the sentences which are actually transformed. The main advantage of this method is that it requires only POS tags; the obtained results are competitive with those presented in previous work, and in contrast to these methods, neither parsing or other type of deep syntactic analysis is required, nor a bilingual word alignment. In addition to applying reorderings as a preprocessing step, translation of word graphs created on the base of these reorderings is investigated, and further improvements of translation quality are achieved.

Translation with scarce bilingual resources

A trade-off between the size of the bilingual training corpus and translation quality has been systematically investigated for three distinct language pairs and different domains. For all tasks, an acceptable translation quality is achieved by training on a very small amount of task-specific parallel text with the help of morpho-syntactic transformations, especially if conventional dictionaries and/or phrasal books are available as additional bilingual knowledge sources. Translation with a system trained only on a dictionary or a phrasal lexicon could be used for applications where only a gist of the translated text is necessary, such as text classification or multilingual information retrieval. Some morpho-syntactic transformations are shown to be particularly useful for scarce training data, such as local POS-based word reorderings and reduction of full forms.

Automatic error analysis and classification

Error analysis of translation output is an important but difficult task. Human error analysis is, like all human evaluations, costly and time-consuming. This work presented a framework for automatic analysis and categorisation of errors in machine translation output. The basic idea is the use of details about the actual errors obtained from standard word error rates WER and PER in combination with linguistic knowledge. The overall goal is to obtain a better overview of the nature of actual translation errors, primarily to obtain ideas about possible improvements in the translation system and to better understand the differences between different systems. There are many possibilities to carry out an automatic error analysis using the proposed methods. In this thesis, five error categories are presented: morphological (inflectional) errors, reordering errors, missing words, extra words and incorrect lexical choice, and novel error rates based on these categories are introduced. The proposed methods can be used for any translation system, the prerequisite being the availability of a morpho-syntactic analyser for the target language.

The results of the proposed automatic methods are compared with the results obtained by human error analysis. Detailed experiments on different types of corpora and various language pairs were carried out and it is shown that the results of automatic error analysis correlate very well with the results of human evaluation.

Furthermore, it is shown that the new error rates are sensitive to differences between translation systems. Some error categories are particularly sensitive to the amount of the training data: mostly incorrect lexical choice, but also reordering errors caused by local differences as well as morphological errors for highly inflected languages like Spanish. The improvements yielded by POS-based reorderings are also investigated, and it is shown that these improvements are reflected well by the proposed metrics. In addition, it is shown that the new measures are sensitive to the differences between distinct translation systems, i.e. that they can show the weak/strong points of particular system.

8 Future directions

From the experience made during this work, these are the main suggestions for future refinements and investigation:

Word graphs based on morpho-syntactic transformations

- Probabilities for graphs created by POS-based reorderings: appropriate methods for defining reordering probabilities.
- Introducing a reordering path for each particular rule: so far, only two possibilities were considered when creating a word graph; the words are reordered or the words are not reordered. However, it would be interesting to investigate the effects of more diverse paths, e.g. for Spanish–English reorderings create separate paths for adjective-noun, for adverb-adjective-noun, for adjective-adjective-noun, etc., and for German–English verb reorderings separate paths for infinitive, for past participle, for finite verbs, for negative particles, etc.
- Training corpus problem: as discussed in Section 4.3, for some tasks the benefit from the word graph is larger if it is translated with the original corpus without reorderings, whereas for some other tasks it is the opposite. Therefore a combination between the two training variants should be investigated, for example the extraction of phrases both from the original and from the reordered corpus.
- Word graphs for morphological transformations: apart from the graphs based on reorderings, i.e. syntactic transformations, it would be interesting to examine translation of word graphs created on the base of different morphological transformations. As already stated in the previous sections, the effectiveness of both morphological and syntactic transformations depend very much on the languages and on the corpus. For example, as seen in Section 5.3, some morphological transformations are very useful for the approach based on small corpora, whereas if the large corpus is available they are not so beneficial. Apart from this, some preliminary experiments have shown that the hierarchical lexicon proposed by [Nießen & Ney 01b], although very effective for small German–English corpora, does not lead to improvements for the Spanish–English language pair. Furthermore, experiments with splitting Spanish verbs into stems and suffixes yielded improvements for a small tourist-oriented corpus [Popović & Ney 04b], but the method had no impact on the TC-STAR corpus. Therefore translation of word graphs containing morphological variations could be useful for different corpora. Such word graphs could be based on split compound words, word stem and suffix, base forms with or without additional information (such as for example the verb POS tags for Serbian described in Section 5.2,

the hierarchical lexicon proposed in [Nießen & Ney 01b]), etc. Appropriate probabilities for those graphs should be investigated as well.

Translation with scarce resources

- Phrasal lexica: a systematic investigation of phrasal lexica as additional training material for scarce training corpora should be conducted along with the use of appropriate morpho-syntactic information.
- Examine other types of morpho-syntactic transformations, including translation of word graphs.
- Translation with large corpora and conventional dictionaries could be carried out for the Serbian–English pair, as well as translation with scarce bilingual corpora for the other language pairs.

Error analysis

- Details about morphological errors: concordance between articles, adjectives and nouns for highly inflected languages; further analysis of errors for particular word classes, like for example identification of verb errors caused by tense or person, adjective errors caused by gender or number; errors caused by suffix/prefix; errors caused by compound words.
- Details about reordering errors: distances between words; other word classes and groups apart from verbs and noun-adjective groups handled in this work.
- Examine other error categories, such as errors produced by particular named entity (NE) tags.
- Systematic comparison of two methods for inflectional and reordering errors: WER and PER details vs. relative differences.
- Use TER instead of WER and compare the results.
- New error rates as evaluation metrics: correlation of ΣER with human judgments.

A Corpora

This appendix summarises information about the different corpora used for the translation experiments described in this work.

A.1 Spanish–English corpora

The Spanish–English corpora used in this work were collected in the framework of the TC-STAR project [tcs 05] and were used in the first and the second TC-STAR evaluation. The training corpus contains more than one million sentences and about 35 million running words of the Spanish and English transcriptions of the European Parliament Plenary Sessions (EPPS). A detailed description of the EPPS data can be found in [Vilar & Matusov$^+$ 05]. The test corpora consist each of about thousand sentences and 25 000 running words. In addition to the EPPS test copora (Test and Test2), for translation from Spanish into English the Spanish Parliament data (ParlEsp) are used in the second TC-STAR evaluation as well as in this work. The number of Out-of-Vocabulary (OOV) words is very low, about 0.5% of the running words for Spanish and 0.2% for English.

The corpus statistics are shown in Table A.1. In order to analyse the effects of data sparseness, two sets of small corpora have been constructed by random selection of sentences from the original corpus. The small corpus referred to as 13k contains about 1% of the original large corpus, and the corpus referred to as 1k contains only thousand sentences. It can be seen how the number of OOV words is increasing with the decrease of the corpus size reaching about 3% for 13k and about 10% for 1k. In the context of translation with sparse training data, two additional bilingual corpora not related to the domain of the test corpus are explored, namely the conventional dictionary and the phrasal lexicon. Statistics about these data are also presented in Table A.1, and rather high OOV rates can be noted: about 15-25%.

A.2 German–English corpora

The EUROPARL corpus used in this work is the German–English part of the European Parliament corpus described in [Koehn & Monz 05] containing transcriptions of German and English European Parliament Plenary Sessions (EPPS). The training part consists of about 700 000 sentences and 15 000 000 running words, and the test has two thousand sentences and about 55 000 running words. The OOV rates are low, 0.7% for German and 0.2% for English.

The corpus statistics can be seen in Table A.2. In order to investigate the effects of scarce resources, a small subset containing about a thousand sentences and 22 000 running words is randomly extracted from the original corpus. The OOV rates for this corpus are significantly

higher, about 16% for German and 10% of English. The statistics for the conventional dictionary and phrasal lexicon used in the experiments with the scarce training corpora are shown in the same table. Since the dictionary has a rather rich vocabulary, there are actually less OOVs than for the small task-specific corpus 1k. The phrasal lexicon has the highest OOV rates, 25.4% for German and 17.8% for English.

For translation from German into English, the VERBMOBIL data are used in addition to the EUROPARL corpus. The VERBMOBIL corpus consists of the dialogues in domain of appointment scheduling, travel planning and hotel reservation. The training part of the corpus used in this work consists of about 58 000 sentences and about 10 000 dictionary entries as a complement. The test corpus is taken from the end-to-end evaluation of the VERBMOBIL project and containes 251 sentences and about 2600 running words. The main differences between this corpus and the EUROPARL corpus are size and domain. The VERBMOBIL corpus is in comparison to the EUROPARL corpus very small, and the domain is rather restricted.

A.3 Serbian–English corpora

The Serbian–English parallel corpus used in the experiments with scarce training corpora is a language course in electronic form. The full corpus (2.6k) is already small, containing less than three thousand sentences and about twenty five thousand running words. In order to investigate extremely sparse training material, a reduced corpus containing only two hundred sentences reffered to as 0.2k has been randomly extracted from the original corpus. For this corpus, a set of short phrases has been investigated as additional bilingual knowledge. The test part of the corpus consists of two hundred sixty sentences and about two thousand running words. In addition to this test, twenty two sentences from the BBC News are used for translation experiments. All statistics related to this corpora can be seen in Table A.4. High OOV rates can be observed for all corpora and both languages, especially for Serbian due to very rich morphology. In the BBC test set almost half of the Serbian running words are OOV.

Table A.1: Corpus statistics for the Spanish–English TC-STAR task (PM = punctuation marks).

			Spanish	English
Training	1.3M	Sentences	1281427	
		Running words+PM	36578514	34918192
		Vocabulary	153124	106496
		Singletons [%]	35.2	36.2
	13k	Sentences	13360	
		Running words+PM	385198	366055
		Vocabulary	22425	16326
		Singletons [%]	47.6	43.7
	1k	Sentences	1113	
		Running words+PM	31022	29497
		Vocabulary	5809	4749
		Singletons [%]	60.8	55.3
	Dictionary	Entries	52566	
		Running words+PM	60964	62011
		Vocabulary	31126	30761
		Singletons [%]	67.7	67.4
	Phrases	Entries	10520	
		Running Words+PM	44289	41850
		Vocabulary	10797	11167
		Singletons [%]	60.9	64.0
Test	test	Sentences	894	1117
		Running words+PM	28591	28492
		Distinct words	4868	4172
		OOVs (1.3M) [%]	0.63	0.37
		OOVs (13k) [%]	3.8	2.8
		OOVs (1k) [%]	11.9	10.5
		OOVs (dict.) [%]	19.7	14.9
		OOVs (phr.) [%]	22.4	17.8
	test2	Sentences	840	1094
		Running words+PM	22774	26917
		Distinct words	4081	3958
		OOVs (1.3M) [%]	0.14	0.25
		OOVs (13k) [%]	2.8	2.6
		OOVs (1k) [%]	10.6	9.4
		OOVs (dict.) [%]	19.1	16.2
		OOVs (phr.) [%]	23.1	19.1
	spParl	Sentences	888	
		Running words+PM	27877	
		Distinct words	4180	
		OOVs (1.3M) [%]	1.1	
		OOVs (13k) [%]	5.0	
		OOVs (1k) [%]	14.6	
		OOVs (dict.) [%]	18.8	
		OOVs (phr.) [%]	22.5	

Table A.2: Corpus statistics for the German–English EUROPARL task (PM = punctuation marks).

			German	English
Training	700k	Sentences	751088	
		Running Words+PM	15257865	16049170
		Vocabulary	205374	74708
		Singletons [%]	49.8	38.3
	1k	Sentences	1072	
		Running Words+PM	21768	22969
		Vocabulary	5082	3995
		Singletons [%]	65.6	56.7
	Dictionary	Entries	292497	
		Running Words+PM	383685	481972
		Vocabulary	138253	82457
		Singletons [%]	60.6	43.8
	Phrases	Entries	10729	
		Running Words+PM	41338	42674
		Vocabulary	13261	11154
		Singletons [%]	70.7	62.7
Test	test	Sentences	2000	
		Running Words+PM	54260	57951
		Distinct Words	9048	6496
		OOVs (700k) [%]	0.7	0.2
		OOVs (1k) [%]	16.4	10.2
		OOVs (dict.) [%]	10.2	4.2
		OOVs (phr.) [%]	25.4	17.8

Table A.3: Corpus statistics for the German–English VERBMOBIL task (PM = punctuation marks).

Training		German	English
	Sentences	71248	
	Running words+PM	554146	583305
	Vocabulary	11367	6871
	Singletons [%]	40.4	36.5
Test	Sentences	251	
	Running words+PM	2628	
	Distinct words	429	
	OOVs [%]	1.7	

Table A.4: Corpus statistics for the Serbian–English task (PM = punctuation marks).

			Serbian	English
Training	2.6k	Sentences	2632	
		Running words+PM	22227	24808
		Vocabulary	4546	2645
		Singletons [%]	60.0	45.8
	0.2k	Sentences	200	
		Running words+PM	1666	1878
		Vocabulary	778	603
		Singletons [%]	79.4	65.5
	Phrases	Entries	351	
		Running words+PM	617	730
		Vocabulary	335	315
		Singletons [%]	71.3	66.3
Test	test	Sentences	260	
		Running words+PM	2100	2336
		Distinct words	891	674
		OOVs (2.6k) [%]	11.7	4.9
		OOVs (0.2k) [%]	35.2	21.8
	BBC	Sentences	22	
		Running words+PM	395	446
		Vocabulary	213	202
		OOVs (2.6k) [%]	44.3	32.1
		OOVs (0.2k) [%]	53.7	43.7

B Evaluation metrics

B.1 Standard evaluation measures

The following evaluation metrics were used in this work for assessment of translation quality:

BLEU (**Bi**l**e**ingual **e**valuation **u**nderstudy):
BLEU [Papineni & Roukos$^+$ 02] is a precision measure based on *n*-gram counts where typically *n*-grams of size $n \in \{1,...,4\}$ are considered. The precision is modified such that multiple references are combined into a single *n*-gram count vector. All hypothesis unigram, bigram, trigram and fourgram counts are collected and divided by their corresponding maximum reference counts. The clipped hypothesis counts are summed and normalised by the total number of hypothesis *n*-grams. The geometric mean of the modified precision scores for a hypothesis is calculated and then multiplied with an exponential brevity penalty factor to penalise too short translations. BLEU is an accuracy measure.

WER (**W**ord **e**rror **r**ate):
The word error rate (WER) is based on the Levenshtein distance [Levenshtein 66]. It is calculated as the minimum number of substitutions, deletions and insertions that have to be performed in order to transform the translation hypothesis into the reference sentence. This is the standard measure for evaluation of automatic speech recognition systems.

PER (**P**osition-independent word **e**rror **r**ate):
The word order of two target sentences can be different even though they are both correct translations. To account for this, the position-independent word error rate PER proposed by [Tillmann & Vogel$^+$ 97] compares the words in the two sentences *without* taking the word order into account. The PER is always lower than or equal to the WER.

TER (**T**ranslation **e**dit/error **r**ate):
TER [Snover & Dorr$^+$ 06] is defined as an extension of WER: in addition to substitutions, deletions and insertions, possible edits include shifts of word sequences. A shift moves a sequence of words within the hypothesis to another location within the hypothesis. Each shift has the same cost regardless of the number of words in the block or distance moved.

CDER (**CD**-distance-based **e**rror **r**ate):
CDER [Leusch & Ueffing$^+$ 06] is based on the block edit distance: the Levenshtein distance is extended by an additional operation called block movement. CDER allows for reordering of blocks at constant cost. The words in the reference have to be covered exactly once, whereas the words in the hypothesis can be covered zero, one, or multiple times. Thus the CDER measure can be seen as recall-oriented (opposite to the BLEU metric which is based on precision).

For all error rates, if multiple references exist, the Levenshtein distance to the closest reference is calculated for each sentence [Nießen & Och$^+$ 00].

B.2 Syntax-oriented evaluation measures

In this section a set of novel simple linguistic-based metrics based on detailed Part-of-Speech (POS) tags is described and evaluated. Although the idea of using the POSBLEU score was mentioned several years ago,[1] none of the experimental results in this direction have been reported yet in the literature. The following metrics are investigated in this work:

- POSBLEU
  The standard BLEU score calculated on the POS tags instead of words;
- POSWER
  The standard word error rate calculated on the POS tags instead of words;
- POSR4GRAM
  Recall measure based on POS-*n*-grams, $n \in \{1,...,4\}$: percentage of *n*-grams in the reference which are also present in the hypothesis;
- POSP4GRAM
  POS-*n*-gram precision: percentage of *n*-grams in the hypothesis which have a counterpart in the reference;
- POSF4GRAM
  POS-*n*-gram-based F-measure: takes into account all *n*-grams which have a counterpart, both in the reference and in the hypothesis.

For the *n*-gram-measures, two types of *n*-gram averaging are investigated: geometric mean and aritmetic mean. Geometric mean is already widely used in the BLEU score, but is also argued not to be optimal because the score becomes equal to zero even if only one of *n*-gram counts is equal to zero.

All evaluation metrics are based on detailed POS tags. The prerequisite is availability of the appropriate POS tagger for the target language. It should be noted that the POS tags cannot be only basic (noun, verb, etc.) but must contain all morpho-syntactic details (e.g. verb tenses, cases, number, gender, etc.).

B.2.1 Evaluation set-up

The results are presented on the English, French, Spanish and German European Parliament texts generated by different translation systems in the framework of the shared task on the 2006 HLT-NAACL Workshop [Koehn & Monz 06] and 2007 ACL Workshop on Statistical Machine Translation [Callison-Burch & Fordyce$^+$ 07]. The objective of the shared task was translation between European languages, namely French, German, Spanish and English. The translation directions were from each of the languages into English, and vice versa. Training and testing

[1] http://www.amtaweb.org/summit/MTSummit/FinalPapers/panel-hovy.pdf

was based on the EUROPARL corpus. In addition, editorials from the Project Syndicate website[2] were collected and used as out-of-domain test data. The corpus statistics of all test data is shown in Tables B.1 and B.2. About fifteen different groups from different institutions participated in the share tasks. Most of the groups use some variant of a phrase-based statistical system. However, SYSTRAN uses a rule-based system which is not task-specific, and in the evaluation 2007 SYSTRAN and NRC submitted joint translation outputs from a hybrid system using both rule-based and statistic approaches. More details about the shared tasks, the data and the participants can be found in [Koehn & Monz 06] and [Callison-Burch & Fordyce$^+$ 07].

Table B.1: Test data for the shared task 2006.

		English	Spanish	French	German
EUROPARL	Sentences	2 000			
	Words	59 307	61 824	66 783	55 533
	Distinct words	6 031	7 719	7 230	8 812
out-of-domain	Sentences	2 000			
	Words	59 307	61 824	66 783	55 533
	Distinct words	6 031	7 719	7 230	8 812

Table B.2: Test data for the shared task 2007.

		English	Spanish	French	German
EUROPARL	Sentences	2 000			
	Words	53 531	55 380	53 981	49 259
	Distinct words	8 558	10 451	10 186	11 106
out-of-domain	Sentences	2 007			
	Words	43 767	50 771	49 820	45 075
	Distinct words	10 002	10 948	11 244	12 322

The new metrics are evaluated on outputs from all translation directions: Spanish, French and German into English and vice versa. Morpho-syntactic annotation of the English and German references and hypotheses is performed using the constraint grammar parser ENGCG [Voutilainen 95] and GERCG [Haapalainen & Majorin 95]. Spanish texts are annotated using the FreeLing analyser [Carreras & Chao$^+$ 04], and French texts usin the Tree Tagger[3]. In this way, all references and hypotheses are provided with detailed POS tags.

Correlation with human judgments: Spearman's rank correlation coefficient ρ is used in the officional evaluation of various metrics in the share tasks. Therefore we calculate the same coefficient in order to measure correlation of the new automatic metrics with adequacy and fluency scores. Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks, and its advantage is that it makes fewer assumptions about the data. The possible values of ρ are between 1 (if all systems are ranked in the same order) and -1 (if all systems are ranked

[2] `http://www.project-syndicate.com/`
[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

in the reverse order). Thus the higher value of ρ for an automatic metric, the more similar it is to the human metric.

B.2.2 Results

For each new metric, the Spearman ρ coefficient with the adequacy and with the fluency score is calculated on the document level. Then the results are summarised by averaging the obtained coefficients over all translation outputs. Both the mean and the median value are calculated. Median value reflects better how often the certain correlation is above or below this value. Average correlations are calculated separately for the 2006 and 2007 data as for the 2007 data the TER and METEOR scores are also available for comparison whereas for the 2006 data only the BLEU score is available.

Average correlations are shown in Table B.3 and Table B.4. In Table B.3 it can be seen that the new measures have a rather high ρ coefficient both with respect to the adequacy and to the fluency score. POSBLEU and POSF4GRAM using the geometric mean have especially high correlation coefficients, followed by POSWER and other *n*-gram-based metrics. Furthermore, all these measures outperform the BLEU score. It can also be seen that the geometric mean averaging of *n*-grams performs slightly better than the arithmetic mean.

Table B.3: Mean and median Spearman correlations for the 2006 data.

2006	adequacy		fluency	
	mean	median	mean	median
BLEU	0.478	0.660	0.495	0.624
POSBLEU	0.546	**0.727**	0.541	**0.645**
POSWER	0.729	**0.708**	0.659	**0.698**
POSF4GRAM gm	0.597	**0.788**	0.566	**0.739**
am	0.577	**0.751**	0.588	**0.740**
POSR4GRAM gm	0.508	0.652	0.525	0.612
am	0.472	0.623	0.494	0.522
POSP4GRAM gm	0.602	**0.712**	0.532	**0.654**
am	0.569	0.610	0.500	0.554

Average correlation coefficients for the 2007 data are presented in Table B.4. Again, POSBLEU, POSWER and POSF4GRAM have high correlation coefficients with both adequacy and fluency, and for this data POSR4GRAM too. These measures outperform the BLEU score as well as the METEOR and the TER metric.

Results of an additional experiment on the 2007 data are shown in Table B.5. This table presents the percentage of the documents on which the new measure outperforms one of the well-known measures, i.e. BLEU, METEOR and TER. It can be seen that in the majority of the cases the POSBLEU metric outperforms all three standard measures, especially with respect to the fluency score. The POSWER and geometric mean POSF4GRAM show a similar behaviour, outperforming the standard measures in the majority of cases but slightly less often than the POSBLEU. The POSR4GRAM score is worse than POSF4GRAM, but still outperforms

Table B.4: Mean and median Spearman correlations for the 2007 data.

2007	adequacy		fluency	
	mean	median	mean	median
BLEU	0.674	0.669	0.584	0.710
METEOR	0.596	0.604	0.538	0.643
TER	0.598	0.68	0.479	0.59
POSBLEU	0.723	**0.742**	0.697	**0.778**
POSWER	0.676	**0.726**	0.651	**0.742**
POSF4GRAM gm	0.576	**0.693**	0.538	**0.736**
am	0.590	**0.693**	0.556	**0.736**
POSR4GRAM gm	0.626	**0.714**	0.618	**0.804**
am	0.600	**0.732**	0.586	**0.827**
POSP4GRAM gm	0.508	0.642	0.438	0.572
am	0.500	0.607	0.429	0.572

Table B.5: Percentage of documents where a particular new measure outperforms the standard measures – test data 2007.

	adequacy			fluency		
	BLEU	METEOR	TER	BLEU	METEOR	TER
POSBLEU	77.3	58.3	75.0	81.8	83.3	83.3
POSWER	63.6	66.7	75.0	63.6	75.0	83.3
POSF4GRAM gm	72.7	58.3	75.0	63.6	75.0	83.3
am	68.2	58.3	75.0	63.6	66.7	66.7
POSR4GRAM gm	63.6	75.0	58.3	68.1	66.7	58.3
am	54.5	75.0	58.3	63.6	58.3	50.0
POSP4GRAM gm	63.6	50.0	75.0	45.4	50.0	58.3
am	54.5	41.7	66.7	36.4	50.0	58.3

the standard measures in 50-70% of cases, and the POSP4GRAM score has the lowest percentage, 30-60%. Additionaly, it can be seen that, again, the geometric mean averaging of the *n*-grams performs better than the artimetic mean.

B.2.3 Conclusions

The use of lingustic-based evaluation measures oriented on the syntactic structure of the sentence is proposed. BLEU and WER are calculated on the detailed Part-of-Speech (POS) tags instead on the words. In addition, precision, recall and F-measure obtained on POS-*n*-grams are investigated. The new measures are tested on the data of the second and third shared task of Statistical Machine Translation Workshop [Koehn & Monz 06, Callison-Burch & Fordyce$^+$ 07]. An extensive analysis of the correlation coefficients between the linguistic-based automatic evaluation metrics and the human judgments is carried out. The obtained results show that the

new metrics correlate very well with human judgments, namely adequacy and fluency scores, as well as that some of the new metrics outperform the standard evaluation measures BLEU, METEOR and TER in the majority of cases. POSBLEU, geometric mean POSF4GRAM and POSWER seem to be especially promising.

Combinations of different evaluation metrics could be investigated in the future, for example the combination of different linguistic-based metric as well as the combination of linguistic metrics with standard metrics.

B.3 Automatic error analysis (Chapter 6)

For the automatic error analysis of translation output following metrics have been used:

INFER (**inf**lectional **e**rror **r**ate):
Number of PER errors caused by wrong choice of the full word form normalised over the (closest) reference length.

RER (**r**eordering **e**rror **r**ate):
Number of WER substitutions and deletions which do not occur as PER reference errors normalised over the (closest) reference length.

MSER (**m**i**ss**ing word **e**rror **r**ate):
Number of WER deletions which are not caused by wrong full form choice normalised over the (closest) reference length.

EXER (**ex**tra word **e**rror **r**ate):
Number of WER insertions which are not caused by wrong full form choice normalised over the (closest) reference length.

LXER (**l**e**x**ical **e**rror **r**ate):
Number of errors caused neither by wrong full form choice nor by deleting or inserting words normalised over the (closest) reference length.

ΣER (**sum** of **e**rror **r**ates):
Sum of all above error categories: always greater than PER, lower than WER and similar to TER.

Two additional measures for estimating inflectional and reordering errors are used:

$\Delta(\textbf{WER}, \textbf{PER})$ – relative difference between WER and PER:
Used for estimation of reordering errors.

$\Delta(\textbf{PER}, \textbf{PER}(b))$ – relative diferrence between PER of full forms and PER of base forms:
Used for estimation of morphological (inflectional) errors.

C Additional experimental results

This Appendix contains experimental results regarding the splitting of German compound words, as well as results of POS-based local reorderings and translation with scarce bilingual data on two additional TC-STAR test sets not included in Chapter 4 and 5.

C.1 Splitting German compound words

This section presents detailed results obtained by the splitting of German compound words on the full EUROPARL training corpus. Two scenarios are investigated: splitting compounds in the original corpus and splitting compounds in the reordered corpus. Table C.1 presents the percentage of sentences which contain split compounds, as well as the percentage of split running words. It can be noted that almost half of sentences in the test corpus are affected by splitting, but only 2.5% of running words are actually split. This is one of the reasons why the improvements achieved by compound splitting are not large. When both long-range verb reorderings and compound splittings are applied, the majority of sentences is transformed.

Table C.1: Percentage of transformed sentences and split words in the German part of the EUROPARL corpus.

		train	test
split	sentences	38.1%	44.6%
	words	4.4%	2.5%
reorder+split	sentences	93.2%	85.7%

The translation results obtained by compound splitting are shown in Table C.2. The baseline system denotes the original system without any transformations, “reorder verbs” stands for the system with the best combination of long-range reorderings described in Section 4.3, and “split compounds” is the system with compound splitting. In the system “reorder+split” both transformations are applied on the source part of the corpus.

The results for the separated test sets are shown in Table C.3. It can be seen that compound splitting results in improvements both for the set of split sentences and for the rest. Contrary to reorderings, improvements by split are even larger for the rest of sentences than for the actually transformed sentences.

POSBLEU scores are calculated for each of the four German-to-English systems and presented in Table C.4. Reordering of verbs, as already seen in Section 4.3, significantly improves this score. However, improvements obtained by compound splitting are rather small. This could

be expected, because compound splitting has not much influence on the syntactic structure of the sentence but on the lexical aspects.

Table C.2: Translation results [%] for German→English translation of the EUROPARL corpus: verb reorderings and compound splitting.

700k	BLEU	TER	WER	PER	CDER
baseline: monotone search	24.4	61.3	66.9	45.9	55.6
reorder verbs	25.6	60.2	65.4	45.7	54.4
split compounds	24.8	61.0	66.6	45.5	55.4
reorder+split	25.6	59.8	65.1	45.2	54.4

Table C.3: Separated translation results [%] for German→English translation of the EUROPARL corpus: transformed senteces and the rest.

		BLEU	TER	WER	PER	CDER
split	baseline: monotone search	24.2	61.9	67.7	46.2	56.4
	split compounds	24.4	61.8	67.6	46.0	56.2
not split	baseline: monotone search	23.4	62.5	67.3	48.1	56.5
	split compounds	23.7	62.0	67.0	47.6	56.3
reordered+split	baseline: monotone search	23.6	62.2	68.0	46.3	56.5
	reorder+split	24.6	60.7	66.2	45.7	55.2
not transformed	baseline: monotone search	37.3	49.8	52.5	40.8	44.0
	reorder+split	38.9	48.0	50.9	39.6	43.8

Table C.4: POSBLEU scores [%] for German→English translation of the EUROPARL corpus: verb reorderings and compound splitting.

	German→English
baseline: monotone search	36.3
reorder verbs	38.2
split compounds	36.5
reorder+split	38.6

Table C.5 presents the POSBLEU scores for the separated test sets. Compound splitting yields small improvements both for the transformed sentences and for the rest. When verb reorderings are also applied, improvements on the transformed set are large, and somewhat smaller for the rest.

Table C.5: Separated POSBLEU scores [%] for German→English translation of the EUROPARL corpus: transformed sentences and the rest.

		German→English
split	baseline: monotone search	37.4
	split compounds	37.5
not split	baseline: monotone search	35.0
	split compounds	35.2
reordered+split	baseline: monotone search	35.7
	reorder+split	38.0
not transformed	baseline: monotone search	44.9
	reorder+split	46.6

C.2 Local POS-based word reorderings and translation with scarce resources on additional Spanish–English test sets

This section presents the results of local reorderings as well as results for translation with scarce bilingual training corpora for the test corpora used in the first TC-STAR evaluation. In addition, for the translation from Spanish into English results for the Spanish Parliament corpus used as additional material in the second TC-STAR evaluation are presented. All results for these corpora show the same trends to those obtained on the test corpus from the second evaluation presented in Sections 4.3 and 5.3.

Table C.6: Percentage of reordered sentences in additional Spanish–English test corpora: local reorderings of adjectives.

	train	test2	spParl
Spanish	62.5%	60.8%	55.7%
English	68.7 %	64.2%	/

Table C.7: Translation results [%] for the additional test corpora: Spanish→English, local reordering of adjectives.

Spanish→English		BLEU	TER	WER	PER	CDER
test2	baseline: monotone search	55.2	32.3	34.3	25.3	31.3
	reorder adjectives	55.7	32.1	33.9	25.3	30.8
spParl	baseline: monotone search	41.2	43.9	47.0	33.1	42.0
	reorder adjectives	41.7	43.7	46.7	33.1	41.6

Table C.8: Separated translation results [%] for the additional test corpora: Spanish→English, reordered sentences and the rest

Spanish→English			BLEU	TER	WER	PER	CDER
test2	reordered	baseline: monotone search	55.3	32.6	34.7	25.0	31.8
		reorder adjectives	55.9	32.3	34.2	25.0	31.2
	not reordered	baseline: monotone search	54.9	31.4	32.8	26.4	29.9
		reorder adjectives	54.8	31.5	32.8	26.3	29.8
spParl	reordered	baseline: monotone search	41.4	44.5	47.7	32.8	42.5
		reorder adjectives	42.2	44.0	47.1	32.8	41.8
	not reordered	baseline: monotone search	40.5	42.6	44.9	34.1	40.3
		reorder adjectives	40.4	42.6	44.9	34.2	40.3

Table C.9: Translation results [%] for the additional test corpora: English→Spanish, local reordering of adjectives.

English→Spanish		BLEU	TER	WER	PER	CDER
test2	baseline: monotone search	48.2	38.9	41.0	31.3	37.6
	reorder adjectives	48.6	38.5	40.6	31.0	37.3

Table C.10: Separated translation results [%] for the additional test corpora: English→Spanish, reordered sentences and the rest.

English→Spanish			BLEU	TER	WER	PER	CDER
test2	reordered	baseline: monotone search	48.1	39.2	41.4	31.0	38.1
		reorder adjectives	48.7	38.7	41.0	30.7	37.6
	not reordered	baseline: monotone search	48.4	38.1	39.4	32.0	36.1
		reorder adjectives	48.4	38.0	39.4	31.9	36.3

Table C.11: POSBLEU scores [%] for both translation directions on the additional Spanish–English test corpora: local reorderings of adjectives.

	Spanish→English		English→Spanish
	test2	spParl	test2
baseline: monotone search	69.7	61.5	57.4
reorder adjectives	70.6	62.2	58.1

Table C.12: Separated POSBLEU scores [%] for both translation directions on the additional Spanish–English test corpora: reordered sentences and the rest.

		Spanish→English		English→Spanish
		test2	spParl	test2
reordered	baseline: monotone search	69.2	61.8	57.3
	reorder adjectives	70.4	62.8	58.3
rest	baseline: monotone search	71.3	60.7	57.5
	reorder adjectives	71.1	60.6	57.4

Table C.13: Translation of word graphs: results [%] on the additional test corpora – Spanish→English.

Spanish→English		BLEU	TER	WER	PER	CDER
test2	baseline: monotone search	55.2	32.3	34.3	**25.3**	31.3
	+graph	**56.1**	**31.7**	**33.7**	**25.3**	**30.8**
	reorder adjectives	55.7	32.1	33.9	**25.3**	**30.8**
	+graph	55.7	32.0	33.9	25.4	**30.8**
spParl	baseline: monotone search	41.2	43.9	47.0	33.1	42.0
	+graph	**42.8**	**42.7**	**45.9**	**32.2**	**40.7**
	reorder adjectives	41.7	43.7	46.7	33.1	41.6
	+graph	42.7	43.0	46.2	**32.2**	40.8

Table C.14: Translation of word graphs: results [%] on the additional test corpora – English→Spanish.

English→Spanish		BLEU	TER	WER	PER	CDER
test2	baseline: monotone search	48.2	38.9	41.0	31.3	37.6
	+graph	**48.8**	38.5	**40.5**	31.2	**37.2**
	reorder adjectives	48.6	38.5	40.6	31.0	37.3
	+graph	48.7	**38.3**	**40.5**	**30.9**	**37.2**

Table C.15: Translation tesults and OOV rates [%] for the additional test corpora: Spanish→English, different sizes of training corpora and appropriate morpho-syntactic transformations.

Spanish→English		BLEU	TER	WER	PER	CDER	OOV rate
test2	dictionary	19.4	58.8	60.4	49.3	56.9	20.7
	+reorder adjectives	20.1	56.7	59.4	47.4	54.8	20.7
	+adjective base	23.8	54.9	56.4	46.8	53.0	17.9
	+ OOV base	24.8	53.8	55.3	44.9	51.9	6.9
	1k	29.6	50.9	52.4	40.8	48.6	10.6
	+dictionary	34.8	46.7	48.4	36.9	45.3	6.8
	+reorder adjectives	39.8	43.4	44.9	35.3	41.4	6.8
	+adjective base	40.1	43.1	44.6	35.0	41.1	5.9
	13k	44.5	39.4	41.3	30.7	37.8	2.8
	+dictionary	46.3	38.2	40.1	29.7	37.5	2.4
	+reorder adjectives	48.9	36.7	38.6	29.2	35.2	2.4
	+adjective base	48.9	36.6	38.3	29.0	35.1	2.2
	1.3M	55.2	32.3	34.3	25.3	31.3	0.14
	reorder adjectives	55.7	32.1	33.9	25.3	30.8	0.14
spParl	dictionary	15.4	65.5	67.8	53.3	62.7	18.8
	+reorder adjectives	16.8	64.4	66.6	53.1	61.4	18.8
	+adjective base	17.4	63.8	66.0	52.5	60.9	17.1
	+ OOV base	18.2	62.9	65.3	51.0	59.8	7.2
	1k	20.4	59.5	61.8	47.6	57.0	14.6
	+dictionary	25.0	55.4	57.8	43.5	53.9	8.8
	+reorder adjectives	28.5	53.5	55.9	42.8	51.4	8.8
	+adjective base	28.5	53.5	55.9	42.7	51.3	8.0
	13k	32.4	50.4	53.3	37.9	48.2	5.0
	+dictionary	33.7	49.1	51.9	37.4	48.0	3.9
	+reorder adjectives	36.2	47.9	50.7	37.1	45.8	3.9
	+adjective base	35.6	48.2	50.9	37.6	46.1	3.8
	1.3M	41.2	43.9	47.0	33.1	42.0	1.1
	+reorder adjectives	41.7	43.7	46.7	33.2	41.6	1.1

Table C.16: Translation results and OOV rates [%] for the additional test corpora: English→Spanish, different sizes of training corpora and appropriate morpho-syntactic transformations.

English→Spanish		BLEU	TER	WER	PER	CDER	OOVs
test2	dictionary	14.1	65.4	67.6	55.9	62.8	16.2
	+reorder adjectives	15.7	64.0	66.3	55.2	61.4	16.2
	1k	24.0	58.8	60.8	47.9	56.0	9.4
	+dictionary	28.4	55.0	57.1	44.2	52.6	4.8
	+reorder adjectives	30.5	53.2	54.0	42.0	50.3	4.8
	13k	36.7	47.9	50.3	38.0	46.0	2.6
	+dictionary	37.7	46.9	49.3	37.1	45.1	1.8
	+reorder adjectives	39.5	45.5	47.8	36.4	44.0	1.8
	1.3M	48.2	38.9	41.0	31.3	37.6	0.25
	reorder adjectives	48.6	38.5	40.6	31.0	37.3	0.25

Table C.17: Translation results [%] for the additional test corpora: Spanish→English, word graphs and scarce training data.

Spanish→English		BLEU	TER	WER	PER	CDER	OOVs
test2	dictionary	19.7	58.8	60.1	48.8	56.9	20.7
	+graph	**22.3**	56.8	58.0	48.3	55.0	
	+reorder adjectives	**22.3**	**56.7**	**57.9**	**48.2**	**54.8**	
	1k	29.6	50.9	52.4	40.8	48.6	10.6
	+graph	32.4	48.6	50.2	39.9	46.4	
	reorder adjectives	**32.9**	**47.9**	**49.4**	39.7	**45.8**	
	+graph	**32.9**	48.0	49.5	**39.6**	45.9	
	13k	44.5	39.4	41.3	30.2	37.8	2.8
	+graph	47.4	37.4	39.2	30.0	35.8	
	reorder adjectives	**47.7**	**37.2**	**39.1**	**29.8**	**35.5**	
	+graph	**47.7**	37.3	**39.1**	**29.8**	35.6	
spParl	dictionary	15.4	65.5	67.8	53.3	62.7	18.8
	+graph	**16.8**	**64.4**	**66.6**	**53.1**	61.5	
	+reorder adjectives	**16.8**	**64.4**	**66.6**	**53.1**	**61.4**	
	1k	20.4	59.5	61.8	47.6	57.0	14.6
	+graph	**22.5**	58.0	**60.1**	47.1	**55.4**	
	reorder adjectives	22.3	**57.8**	**60.1**	**47.0**	**55.4**	
	+graph	22.2	58.0	60.3	**47.0**	**55.4**	
	13k	32.4	50.4	53.3	37.9	48.2	5.0
	+graph	34.6	48.9	51.7	**37.6**	46.7	
	reorder adjectives	**35.0**	**48.5**	**51.2**	**37.6**	**46.5**	
	+graph	**35.0**	48.6	51.3	**37.6**	**46.5**	

Table C.18: Translation results [%] for the additional test corpora: English→Spanish, word graphs and scarce training data.

English→Spanish		BLEU	TER	WER	PER	CDER	OOVs
test2	dictionary	17.0	65.4	66.8	55.0	62.8	16.2
	+graph	**18.9**	**63.6**	**65.2**	**54.0**	**61.2**	
	+reorder adjectives	18.5	64.0	65.5	54.2	61.4	
	1k	24.0	58.8	60.8	47.9	56.0	9.4
	+graph	**26.3**	**56.6**	**58.7**	**46.7**	**54.0**	
	reorder adjectives	25.5	57.1	59.3	46.8	54.3	
	+graph	25.9	56.9	59.0	**46.7**	**54.0**	
	13k	36.7	47.9	50.3	38.0	46.0	2.6
	+graph	**39.0**	**45.9**	**48.3**	**37.0**	**44.3**	
	reorder adjectives	38.3	46.4	48.7	37.3	45.0	
	+graph	38.8	46.0	48.4	37.1	44.5	

Table C.19: Translation results [%] on the additional test corpora: Spanish→English, phrasal lexicon with local reorderings and word graphs.

Spanish→English		BLEU	TER	WER	PER	CDER	OOVs
test2	phrases	22.0	56.9	58.3	47.6	54.2	23.1
	+graph	**23.7**	55.6	56.8	**47.3**	**52.8**	
	reorder adjectives	23.5	**55.4**	**56.7**	**47.3**	52.9	
	+graph	23.3	55.7	57.0	47.4	53.2	
	phrases+dictionary	28.3	51.6	53.3	42.0	49.4	14.7
	+graph	31.3	49.6	51.0	41.3	47.2	
spParl	phrases	16.6	63.5	65.6	52.3	60.5	22.5
	+graph	**19.9**	60.4	**62.5**	**49.5**	**57.2**	
	reorder adjectives	19.6	**60.3**	62.6	**49.5**	57.5	
	+graph	19.6	60.5	62.7	**49.5**	57.6	
	phrases+dictionary	24.5	57.4	60.0	45.3	54.8	13.4
	+graph	26.4	55.9	58.4	45.0	53.2	

Table C.20: Translation results [%] on the additional test corpora: English→Spanish, phrasal lexicon with local reorderings and word graphs.

English→Spanish		BLEU	TER	WER	PER	CDER	OOVs
test2	phrases	17.2	64.9	66.6	54.5	62.5	19.4
	+graph	18.6	**63.5**	**65.2**	**53.9**	**61.1**	
	reorder adjectives	18.2	63.8	65.4	54.1	61.5	
	+graph	**18.7**	**63.5**	**65.2**	**53.9**	61.2	
	phrases+dictionary	22.3	59.7	61.8	49.1	57.8	12.6
	+graph	24.6	57.9	59.9	48.1	55.9	

Bibliography

[Al-Onaizan & Germann+ 00] Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada: Translating with Scarce Resources. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pp. 672–678, Austin, TX, July 2000.

[Babych & Hartley 04] B. Babych, A. Hartley: Extending BLEU MT Evaluation Method with Frequency Weighting. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 04)*, pp. 621–628, Barcelona, Spain, July 2004.

[Banerjee & Lavie 05] S. Banerjee, A. Lavie: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65–72, Ann Arbor, MI, June 2005.

[Bender 02] O. Bender: Untersuchung zur Tagging-Aufgabestellung in der Sprachverarbeitung. Diploma thesis, RWTH Aachen University, Aachen, Germany, October 2002.

[Berger & Brown+ 96] A.L. Berger, P.F. Brown, S.A.D. Pietra, V.J.D. Pietra, J.R. Gillett, A.S. Kehler, R.L. Mercer: Language Translation apparatus and method of using context-based translation models, United States Patent 5510981, April 1996.

[Brants 00] T. Brants: Tnt – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 00)*, pp. 224–231, Seattle, WA, April/May 2000.

[Brown & Cocke+ 90] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin: A Statistical Approach to Machine Translation. *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, June 1990.

[Brown & Della Pietra+ 92] P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, J.D. Lafferty, R.L. Mercer: Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 93)*, pp. 83–100, Kyoto, Japan, July 1992.

[Brown & Della Pietra+ 93] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, M.J. Goldsmith, J. Hajic, R.L. Mercer, S. Mohanty: But Dictionaries Are Data Too. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 202–205, Plainsboro, NJ, March 1993.

[Callison-Burch & Fordyce+ 07] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, J. Schroeder: (Meta-)Evaluation of Machine Translation. In *Proceedings of the 2nd ACL 07 Workshop on Statistical Machine Translation (WMT 07)*, pp. 136–158, Prague, Czech Republic, June 2007.

[Callison-Burch & Osborne 03] C. Callison-Burch, M. Osborne: Co-training for statistical machine translation. In *Proceedings of the 6th Annual CLUK Research Colloquium*, Edinburgh, UK, January 2003.

[Carreras & Chao$^+$ 04] X. Carreras, I. Chao, L. Padró, M. Padró: FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 04)*, pp. 239–242, Lisbon, Portugal, May 2004.

[Collins & Koehn$^+$ 05] M. Collins, P. Koehn, I. Kučerová: Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05)*, pp. 531–540, Ann Arbor, MI, June 2005.

[Costa-jussà & Fonollosa 06] M.R. Costa-jussà, J.A.R. Fonollosa: Statistical Machine Reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 06)*, pp. 70–76, Sydney, Australia, July 2006.

[Crego & de Gispert$^+$ 06] J.M. Crego, A. de Gispert, P. Lambert, M.R. Costa-jussà, M. Khalilov, R. Banchs, J.B. Mariño, J.A.R. Fonollosa: N-gram-based SMT System Enhanced with Reordering Patterns. In *Proceedings of the 1st NAACL 06 Workshop on Statistical Machine Translation (WMT 06)*, pp. 162–165, New York, NY, June 2006.

[Crego & Mariño 06] J.M. Crego, J.B. Mariño: Integration of POS tag-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 (AMTA 06)*, pp. 29–36, Boston, MA, August 2006.

[Creutz & Lagus 02] M. Creutz, K. Lagus: Unsupervised discovery of morphemes. In *The 40th Annual Meeting of the Association for Computational Linguistics (ACL 02): Proceedings of the Workshop on Morphological and Phonological Learning*, pp. 21–30, Philadelphia, PA, July 2002.

[de Gispert & Gupta$^+$ 06] A. de Gispert, D. Gupta, M. Popović, P. Lambert, J.B. Mariño, M. Federico, H. Ney, R. Banchs: Improving Statistical Word Alignments with Morpho-syntactic Transformations. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL)*, pp. 368–379, Turku, Finland, August 2006. Lecture Notes in Computer Science, Springer Verlag.

[de Gispert & Mariño$^+$ 05] A. de Gispert, J.B. Mariño, J.M. Crego: Improving Statistical Machine Translation by Classifying and Generalizing Inflected Verb Forms. In *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech 05)*, pp. 3185–3188, Lisbon, Portugal, September 2005.

[Doddington 02] G. Doddington: Automatic Evaluation of Machine Tanslation Quality using n-gram Co-occurrence Statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 128–132, San Diego, March 2002.

[Fraser & Marcu 06] A. Fraser, D. Marcu: Measuring Word Alignment Quality for Statistical Machine Translation. Technical report, ISI-University of Southern California, May 2006.

[Giménez & Amigó 06] J. Giménez, E. Amigó: IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pp. 685–690, Genoa, Italy, May 2006.

[Goldsmith 01] J. Goldsmith: Unsupervised Learning of the Morphology of a Natural language. *Computational Linguistics*, Vol. 27, No. 2, pp. 153–198, June 2001.

[Goldwater & McClosky 05] S. Goldwater, D. McClosky: Improving stastistical machine translation through morphological analysis. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 05)*, pp. 676–683, Vancouver, Canada, October 2005.

[Haapalainen & Majorin 95] M. Haapalainen, A. Majorin: GERTWOL und Morphologische Disambiguierung für das Deutsche. http://www.lingsoft.fi/doc/gercg/NODALIDA-poster.html, 1995.

[Kanthak & Vilar$^+$ 05] S. Kanthak, D. Vilar, E. Matusov, R. Zens, H. Ney: Novel Reordering Approaches in Phrase-Based Statistical Machine Translation. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 167–174, Ann Arbor, MI, June 2005.

[Kirchhoff & Yang$^+$ 06] K. Kirchhoff, M. Yang, K. Duh: Statistical Machine Translation of Parliamentary Proceedings Using Morpho-Syntactic Knowledge. In *Proceedings of the TC-Star Workshop on Speech-to-Speech Translation*, pp. 57–62, Barcelona, Spain, June 2006.

[Kneser & Ney 95] R. Kneser, H. Ney: Improved backing-off for M-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 95)*, Vol. 1, pp. 181–184, Detroit, MI, May 1995.

[Knight 99] K. Knight: Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, Vol. 25, No. 4, pp. 607–615, December 1999.

[Koehn & Knight 03] P. Koehn, K. Knight: Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 03)*, pp. 347–354, Budapest, Hungary, April 2003.

[Koehn & Monz 05] P. Koehn, C. Monz: Shared task: statistical machine translation between European languages. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 119–124, Ann Arbor, MI, June 2005.

[Koehn & Monz 06] P. Koehn, C. Monz: Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the 1st NAACL 06 Workshop on Statistical Machine Translation (WMT 06)*, pp. 102–121, New York, NY, June 2006.

[Larson 01] M. Larson: Sub-Word-Based Language Models for Speech Recognition: Implications for Spoken Document Retrieval. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, CMU, Pittsburgh, PA, June 2001.

[Larson & Willett$^+$ 00] M. Larson, D. Willett, J. Köhler, G. Rigoll: Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 00)*, Vol. 3, pp. 945–948, Beijing, China, February 2000.

[Lee 04] Y. Lee: Morphological analysis for statistical machine translation. In *Proceedings of the Meeting of the North American chapter of the Association for Computational Linguistics (HLT-NAACL 04)*, pp. 57–60, Boston, MA, May 2004.

[Lee & Ge 06] Y.S. Lee, N. Ge: Local Reordering in Statistical Machine Translation. In *Proceedings of the TC-Star Workshop on Speech-to-Speech Translation*, pp. 93–98, Barcelona, Spain, June 2006.

[Leusch & Ueffing$^+$ 05] G. Leusch, N. Ueffing, D. Vilar, H. Ney: Preprocessing and Normalization for Automatic Evaluation of Machine Translation. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 17–24, Ann Arbor, MI, June 2005.

[Leusch & Ueffing$^+$ 06] G. Leusch, N. Ueffing, H. Ney: CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 06)*, pp. 241–248, Trento, Italy, April 2006.

[Levenshtein 66] V.I. Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710, February 1966.

[Liu & Gildea 05] D. Liu, D. Gildea: Syntactic features for evaluation of machine translation. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 25–32, Ann Arbor, MI, June 2005.

[Llitjós & Carbonell$^+$ 05] A.F. Llitjós, J.G. Carbonell, A. Lavie: A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 05)*, pp. 87–96, Budapest, Hungary, May 2005.

[Matusov & Leusch$^+$ 05] E. Matusov, G. Leusch, O. Bender, H. Ney: Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 05)*, pp. 148–154, Pittsburgh, PA, October 2005.

[Matusov & Zens$^+$ 06] E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popović, H. Ney: The RWTH Machine Translation System. In *Proceedings of the TC-Star Workshop on Speech-to-Speech Translation*, pp. 31–36, Barcelona, Spain, June 2006.

[Melamed & Green$^+$ 03] I.D. Melamed, R. Green, J.P. Turian: Precision and Recall of Machine Translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL 03)*, pp. 61–63, Edmonton, Canada, May/June 2003.

[Ney & Popović$^+$ 04] H. Ney, M. Popović, D. Sündermann: Error Measures and Bayes Decision Rules Revisited with Applications to POS Tagging. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 04)*, pp. 270–276, Barcelona, Spain, July 2004.

[Nießen 02] S. Nießen: *Improving Statistical Machine Translaton using Morpho-syntactic Information*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, December 2002.

[Nießen & Ney 00] S. Nießen, H. Ney: Improving SMT quality with morpho-syntactic analysis. In *Proceedings of the 18th International Conference on Computational Linguistics (CoLing 00)*, pp. 1081–1085, Saarbrücken, Germany, July 2000.

[Nießen & Ney 01a] S. Nießen, H. Ney: Morpho-syntactic analysis for Reordering in Statistical Machine Translation. In *Proceedings of the MT Summit VIII*, pp. 247–252, Santiago de Compostela, Spain, September 2001.

[Nießen & Ney 01b] S. Nießen, H. Ney: Toward hierarchical models for statistical machine translation of inflected languages. In *The 39th Annual Meeting of the Association for Computational Linguistics (ACL 01): Proceedings of the Workshop on Data-Driven Machine Translation*, pp. 47–54, Toulouse, France, July 2001.

[Nießen & Ney 04] S. Nießen, H. Ney: Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, Vol. 30, No. 2, pp. 181–204, June 2004.

[Nießen & Och$^+$ 00] S. Nießen, F.J. Och, G. Leusch, H. Ney: An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 00)*, pp. 39–45, Athens, Greece, May 2000.

[Och 03] F.J. Och: Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL 03)*, pp. 160–167, Sapporo, Japan, July 2003.

[Och & Ney 02] F.J. Och, H. Ney: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pp. 295–302, Philadelphia, PA, July 2002.

[Och & Ney 03] F.J. Och, H. Ney: A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, March 2003.

[Och & Ney 04] F.J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417–449, December 2004.

[Owczarzak & van Genabith$^+$ 07] K. Owczarzak, J. van Genabith, A. Way: Dependency-based automatic evaluation for machine translation. In *Proceedings of the NAACL-HLT 2007 Workshop on Syntax and Structure in Statistical Translation*, pp. 80–87, Rochester, NY, April 2007.

[Papineni & Roukos$^+$ 02] K. Papineni, S. Roukos, T. Ward, W.J. Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pp. 311–318, Philadelphia, PA, July 2002.

[Popović & de Gispert$^+$ 06] M. Popović, A. de Gispert, D. Gupta, P. Lambert, H. Ney, J.B. Mariño, M. Federico, R. Banchs: Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 1st NAACL 06 Workshop on Statistical Machine Translation (WMT 06)*, pp. 1–6, New York, NY, June 2006.

[Popović & Jovičić$^+$ 04] M. Popović, S. Jovičić, Z. Šarić: Statistical Machine Translation of Serbian-English. In *Proceedings of the International Workshop on Speech and Computer (SPECOM)*, pp. 410–414, St. Petersburg, Russia, September 2004.

[Popović & Ney 04a] M. Popović, H. Ney: Improving Word Alignment Quality using Morpho-Syntactic Information. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing 04)*, pp. 310–314, Geneva, Switzerland, August 2004.

[Popović & Ney 04b] M. Popović, H. Ney: Towards the Use of Word Stems & Suffixes for Statistical Machine Translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 04)*, pp. 1585–1588, Lissabon, Portugal, May 2004.

[Popović & Ney 05] M. Popović, H. Ney: Exploiting Phrasal Lexica and Additional Morpho-syntactic Language Resources for Statistical Machine Translation with Scarce Training Data. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 05)*, pp. 212–218, Budapest, Hungary, May 2005.

[Popović & Ney 06a] M. Popović, H. Ney: Error Analysis of Verb Inflections in Spanish Translation Output. In *Proceedings of the TC-Star Workshop on Speech-to-Speech Translation*, pp. 99–103, Barcelona, Spain, June 2006.

[Popović & Ney 06b] M. Popović, H. Ney: POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pp. 1278–1283, Genoa, Italy, May 2006.

[Popović & Ney 06c] M. Popović, H. Ney: Statistical Machine Translation with a Small Amount of Bilingual Training Data. In *Proceedings of the LREC 06 Workshop on Strategies for Developing Machine Translation for Minority Languages*, pp. 25–29, Genoa, Italy, May 2006.

[Popović & Ney 07] M. Popović, H. Ney: Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the 2nd ACL 07 Workshop on Statistical Machine Translation (WMT 07)*, pp. 48–55, Prague, Czech Republic, June 2007.

[Popović & Ney 09] M. Popović, H. Ney: Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the 4th EACL 09 Workshop on Statistical Machine Translation (WMT 09)*, pp. 29–32, Athens, Greece, March 2009.

[Popović & Stein$^+$ 06] M. Popović, D. Stein, H. Ney: Statistical Machine Translation of German Compound Words. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL)*, pp. 616–624, Turku, Finland, August 2006. Lecture Notes in Computer Science, Springer Verlag.

[Popović & Vilar$^+$ 05] M. Popović, D. Vilar, H. Ney, S. Jovičić, Z. Šarić: Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian–English Statistical Machine Translation. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 05): Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 41–48, Ann Arbor, MI, June 2005.

[Ratnaparkhi 96] A. Ratnaparkhi: A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP 96)*, pp. 133–142, Pensylvania, May 1996.

[Rottmann & Vogel 07] K. Rottmann, S. Vogel: 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 (TMI 07)*, pp. 171–180, Skövde, Sweden, September 2007.

[Siivola & Hirsimäki$^+$ 03] V. Siivola, T. Hirsimäki, M. Creutz, M. Kurimo: Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech 03)*, pp. 2293–2296, Geneva, Switzerland, September 2003.

[Snover & Dorr+ 06] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul: A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 06)*, pp. 223–231, Boston, MA, August 2006.

[Stolcke 02] A. Stolcke: SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 02)*, Vol. 2, pp. 901–904, Denver, CO, September 2002.

[tcs 05] TC-STAR - Technology and Corpora for Speech to Speech Translation, 2005. Integrated project TCSTAR (IST-2002-FP6-506738) funded by the European Commission. `http://www.tc-star.org/`.

[Tillmann & Vogel+ 97] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, H. Sawaf: Accelerated DP Based Search for Statistical Translation. In *Proceedings of the 2nd European Conference on Speech Communication and Technology (EuroSpeech 97)*, pp. 2667–2670, Rhodes, Greece, September 1997.

[Toutanova & Ilhan+ 02] K. Toutanova, H.T. Ilhan, C. Manning: Extensions to HMM-based statistical word alignment models. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 02)*, pp. 87–94, Philadelphia, PA, July 2002.

[Toutanova & Manning 00] K. Toutanova, C. Manning: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pp. 63–70, Hong Kong, October 2000.

[Turian & Shen+ 03] J. Turian, L. Shen, I.D. Melamed: Evaluation of Machine Translation and its Evaluation. In *Proceedings of the MT Summit IX*, pp. 23–28, New Orleans, LA, September 2003.

[Ueffing & Ney 03] N. Ueffing, H. Ney: Using POS Information for Statistical Machine Translation into Morphologically Rich Languages. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 03)*, pp. 347–354, Budapest, Hungary, April 2003.

[Vilar & Matusov+ 05] D. Vilar, E. Matusov, S. Hasan, R. Zens, H. Ney: Statistical Machine Translation of European Parliamentary Speeches. In *Proceedings of the MT Summit X*, pp. 259–266, Phuket, Thailand, September 2005.

[Vilar & Popović+ 06] D. Vilar, M. Popović, H. Ney: AER: Do we need to "improve" our alignments? In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 06)*, pp. 205–212, Kyoto, Japan, November 2006.

[Vilar & Xu+ 06] D. Vilar, J. Xu, L.F. D'Haro, H. Ney: Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pp. 697–702, Genoa, Italy, May 2006.

[Vogel & Monson 04] S. Vogel, C. Monson: Augmenting Manual Dictionaries for Statistical Machine Translation Systems. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 04)*, pp. 1589–1592, Lissabon, Portugal, May 2004.

[Voutilainen 95] A. Voutilainen: ENGCG - Constraint Grammar Parser of English. http://www2.lingsoft.fi/doc/engcg/intro/, 1995.

[Wang & Collins+ 07] C. Wang, M. Collins, P. Koehn: Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning (EMNLP/CoNLL 07)*, pp. 737–745, Prague, Czech Republic, June 2007.

[Wu 97] D. Wu: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, Vol. 23, No. 3, pp. 377–403, September 1997.

[Zens & Bender+ 05] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, H. Ney: The RWTH Phrase-based Statistical Machine Translation System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 05)*, pp. 155–162, Pittsburgh, PA, October 2005.

[Zens & Och+ 02] R. Zens, F.J. Och, H. Ney: Phrase-Based Statistical Machine Translation. In M. Jarke, J. Koehler, G. Lakemeyer, editors, *25th German Conference on Artificial Intelligence (KI2002)*, Vol. 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, pp. 18–32, Aachen, Germany, September 2002. Springer Verlag.

[Zhang & Zens+ 07] Y. Zhang, R. Zens, H. Ney: Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Proceedings of the Meeting of the North American chapter of the Association for Computational Linguistics (HLT-NAACL 07)*, pp. 1–8, Rochester, NY, April 2007.

Lebenslauf - Curriculum Vitae

Persönliche Angaben

Name:	Maja Popović
Adresse:	Franzstrasse 34, 52064 Aachen
Geburtsdatum:	16. Juni 1971
Geburtsort:	Belgrad, Serbien
Staatsangehörigkeit:	serbisch

Schulbildung

1977 - 1985:	Grundschule “Karađorđe”, Belgrad
1977 - 1984:	Musikgrundschule “Petar Konjović”, Belgrad
1985 - 1987:	XII Gymnasium, Belgrad
1987 - 1989:	Mathematisches Gymnasium, Belgrad (Abitur)

Studium

01.10.1989 - 07.07.1995:	Studium der Elektrotechnik an der Fakultät für Elektrotechnik, Universität von Belgrad
07.07.1995:	Abschluss als Diplom-Ingenieurin der Elektrotechnik
01.10.1995 - 30.06.1998:	Postdiplomstudium der Elektrotechnik an der Fakultät für Elektrotechnik, Universität von Belgrad
30.06.1998:	“Nachdiplom” in Elektrotechnik - Sprachanalyse

Arbeitstätigkeiten

Oktober 1995 - Oktober 1998:	Wissenschaftliche Angestellte an der Fakultät für Elektrotechnik, Universität von Belgrad
November 1998 - Februar 2000:	Forschungsingenieur am Institut “Mihajlo Pupin”, Universität von Belgrad
März 2000 - September 2001:	Wissenschaftliche Angestellte am Institut “IDIAP”, EPFL - Lausanne
seit Februar 2002:	Wissenschaftliche Angestellte am Lehrstuhl für Informatik 6 der RWTH Aachen

Sprachen:

Serbisch	Muttersprache
English	Fließend im Wort und Schrift
Deutsch	Fließend im Wort und Schrift
Spanisch	Fließend im Wort und Schrift
Französisch	Wort und Schrift