



UPPSALA  
UNIVERSITET

Department of Linguistics and Philology  
Språkteknologiprogrammet  
(Language Technology Programme)  
Master's thesis in Computational Linguistics

8th June 2005

# The Impact of Lemmatization in Word Alignment

Peter Strömbäck

Supervisors:  
Anna Sâgvall Hein, Uppsala Universitet  
Eva Pettersson, Uppsala Universitet

## Abstract

The focus of this thesis is on examining whether word alignment results can be improved in precision and recall through lemmatization, and extraction of lemma dictionaries from the resulting links. Lemmas are extracted from existing lexical resources in order to replace word forms in two parallel corpora documents, one featuring the language pair English-Swedish and the other the language pair Swedish-English.

The parallel corpora, consisting of a technical Scania manual and a Saul Bellow novel, originate from PLUG (Sågvald Hein 2002) project and were originally aligned and evaluated by Jörg Tiedemann (2003).

By utilizing a Perl script, four lemmatized documents are created. These are aligned by a word aligner constructed by Tiedemann (2003), the *Clue aligner*, which is also used to align word form versions of the same texts. The results of the alignment of the lemmatized corpora and the word form corpora are evaluated automatically against a reference alignment and compared.

The link results derived from lemmatized corpora yields improvement in recall, and in one case precision, compared to the word form link results.

Furthermore, lemma extensions are included in aligning experiments in order to extract two lemma form translation dictionaries as a possible lexical resource addition to the MATS system.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>List of Tables</b>	<b>iv</b>
<b>Preface</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Aims</b>	<b>2</b>
<b>3 Outline</b>	<b>3</b>
<b>4 Background</b>	<b>4</b>
4.1 Introduction . . . . .	4
4.2 Alignment . . . . .	4
4.2.1 Sentence alignment . . . . .	4
4.2.2 Word alignment . . . . .	5
4.3 Parallel corpora . . . . .	5
4.4 Methods for word alignment . . . . .	6
4.4.1 The association approach vs. the estimation approach . . . . .	6
4.4.2 The <i>Clue aligner</i> . . . . .	7
4.5 Applications of word alignment . . . . .	8
4.6 Evaluation metrics . . . . .	10
4.7 Lemmatization . . . . .	10
4.7.1 Definition of a lemma . . . . .	10
4.7.2 NLP systems for morphological analysis . . . . .	10
<b>5 Method</b>	<b>12</b>
5.1 Deriving lemma forms . . . . .	12
5.1.1 Deriving Swedish lemma forms . . . . .	12
5.1.2 Deriving English lemma forms . . . . .	13
5.2 Construction of Perl script . . . . .	14
5.3 Compound analyzing . . . . .	15
5.4 Aligning with the <i>Clue aligner</i> . . . . .	15
5.5 Evaluation of word links . . . . .	15
5.6 Application of extracted data . . . . .	15

<b>6</b>	<b>Pre-processing and lemmatizing</b>	<b>17</b>
6.1	Corpora background . . . . .	17
6.2	Pre-processing and format of corpora . . . . .	17
6.3	Extraction of lemmas . . . . .	18
6.4	Replacing word forms with lemmas . . . . .	19
<b>7</b>	<b>Results and evaluation of word links</b>	<b>21</b>
<b>8</b>	<b>Application of extracted data</b>	<b>23</b>
8.1	Extraction of lemma dictionaries . . . . .	23
<b>9</b>	<b>Discussion</b>	<b>26</b>
	<b>Bibliography</b>	<b>27</b>
	<b>Appendix</b>	<b>30</b>

# List of Tables

5.1	Lexicon file example . . . . .	13
6.1	Corpora . . . . .	17
6.2	Lexicon files for Swedish . . . . .	19
6.3	Lexicon files for English . . . . .	19
6.4	Success rate for lemmatized corpora . . . . .	20
6.5	Type-token ratio for lemmatized corpora . . . . .	20
7.1	Linking results for parallel Bellow corpora . . . . .	21
7.2	Linking results for parallel Scania95 corpora . . . . .	22
8.1	Linking results for parallel Bellow corpora with lemma extensions . . .	24
8.2	Linking results for parallel Scania95 corpora with lemma extensions . .	24

# Preface

I would like to give a big thank you to my supervisors: head supervisor Anna Sågvall Hein for feedback on the overall planning and direction of the thesis and for coming up with the idea, Eva Pettersson who unselfishly offered invaluable help with and feedback on daily work, especially programming. Thanks to Jörg Tiedemann for technical support on word linking. Also, a hug and thank you to my classmates who struggled in front of nearby computers for feedback and great fun.

# 1 Introduction

Re-use of translations for statistical machine translation is increasing in popularity within international computer linguistics research. A central issue is deriving lexical data from parallel corpora, the results of which can be of use in machine translation systems, translation support and bi- and multilingual lexicons.

In *Recycling Translations* (Tiedemann 2003), a combination of translational *clues* were applied to word alignment. A series of experiments were carried out on corpora from the PLUG (Sågvall Hein 2002) project. In the experiments, translation probabilities from the statistical alignment tool *GIZA++* (Och 2002) were included, and this combination of statistical and linguistic information resulted in an impressive performance rate of above 80 % in precision and recall (Tiedemann 2003).

This thesis relates to the previous work done by Tiedemann, with a view to improve word linking results by using lemmatized corpora. The produced alignments are evaluated and then extracted for further use as lemma lexicons.

## 2 Aims

The aim of this thesis is to examine whether word alignment results can be improved using lemmatized corpora, building on a hypothesis that lemmatized parallel corpora will lead to improved link results, as word forms from the same paradigm can be generalized to lemmas, resulting in fewer and more frequent words.

Moreover, the possibilities of extracting the produced alignments for use in a Natural Language Processing (NLP) lexicon, more specifically as a lexical resource for the machine translation system MATS in the form of a lemma lexicon, are examined. Lemma lexicons are usually produced from word form links, for which lemmas are assigned. Aligning already lemmatized corpora can save time and energy when creating dictionaries, as the result will automatically be in lemma form.

## 3 Outline

This thesis includes six chapters providing background and overview of the subject, description of applied tools and methods, and evaluation of the results.

Chapter 4 details different word and sentence alignment approaches. The *Clue aligner* and the theory behind it is briefly described. Information is also given on applications of word alignment, parallel corpora, methods for evaluation of NLP systems, and lemmatization.

Chapter 5 clarifies the goals of the thesis, and states important decisions taken. Selection and extraction of lemmas from existing lexical resources yields lemmas. These lemmas are used to replace word forms in bitexts by running a Perl script, the implementation of which is also described. Furthermore, tools for alignment, evaluation and creation of applications are described.

Chapter 6 describes pre-processing of parallel corpora and lemmatization of word forms, along with measures giving the percentage of word forms that were lemmatized.

Chapter 7 presents the results of the word alignments and an evaluation of each alignment.

Chapter 8 contains a description and discussion of the extracted lemma lexicons.

Chapter 9 is a concluding discussion of word alignment results, extracted lemma dictionaries and, briefly, future work.

## 4 Background

### 4.1 Introduction

This chapter introduces concepts and strategies that are relevant to my thesis. The theory and development of word and sentence alignment and statistical machine translation are briefly outlined.

Sentence and word alignment are techniques for mapping sentences and words which are translations of each other in bilingual parallel texts. When aligning, the source language text is split into segments that correspond to segments in the target language. The most common text units are sentences and paragraphs. Links are established between the corresponding segments, and when this is carried out at sentence level it is called *sentence alignment*.

*Word alignment* produces links between the many corresponding words and phrases in the aligned sentences in a parallel corpus. Most methods of word alignment presume that the texts are already sentence-aligned; a pair of sentences that are translations of each other are likely to contain words that are directly or indirectly translations of each other.

### 4.2 Alignment

#### 4.2.1 Sentence alignment

Sentence alignment assumes that the information at sentence level is expressed in the very same order in the source text as in the translation. Thus, the alignment is seen as *monotonic*; without crossing links (Tiedemann 2003). The alignment is not always 1:1; it can also be 2:1, 2:3, 1:0 or 1:2 and so on. For example, one sentence in the source language may correspond to two sentences in the target language, or no sentence at all.

Sentence alignment has been approached in different ways. One of the first and most documented sentence alignment algorithms was introduced by Gale and Church (1991). In short, given a number of possible alignments, it is expected to find the maximum likelihood alignment of sentences (Songlin Piao 2001). Gale and Church's algorithm is based on sentence length, since the authors found a high correlation between character length of corresponding sentences in many language pairs (Tiedemann 2003).

Kay and Roscheisen (1993) approached the problem from a different angle. They created an algorithm using lexical information, where sentence translations are detected based on word translations contained by a pair of candidate sentences, i.e.

selected words with similar distribution. The algorithm performs well but is apparently slower than Gale and Church's algorithm.

A combination of sentence length and lexical information has been proposed by for example Simard et. al. (1992). In addition to Gale and Church's algorithm they used *cognates*, i.e. pairs of tokens from source and target language which are etymologically related, to identify sentence alignments. They achieved only a modest improvement over Gale and Church (Songlin Piao 2001).

## 4.2.2 Word alignment

Aligning words is a more complex process than aligning sentences for several reasons. There are often few one-to-one correspondences between the words in the documents, and a word in the source language may be translated into a number of different words in the target language, depending on the context of word (Songlin Piao 2001). Also, the mappings from word-to-word are often not monotonic, as word order is not identical for most languages. To complicate word alignment further, boundaries of lexical units are harder to detect than sentence boundaries (which are marked by stops etc.) and there is no reliable correlation between character length of words in the languages aligned (Tiedemann 2003). The common goal for word alignment is that all lexical items in a corpus should be aligned, which may lead to "fuzzy" alignments. The difference between a regular and a "fuzzy" alignment is that the source and target units in the latter are different in degrees of specification or semantically overlapping (Merkel et. al 1999a). As an example: a translation from English to Swedish, where the translation of the phrase "came out" is "fuzzy"<sup>1</sup>:

(1) "The spiders *came out* from behind their pictures."

(2) "Spindlarna *hade vågat sig ut* från sina tillhåll."

(Merkel et. al 1999a:157).

## 4.3 Parallel corpora

Corpora is a foundation block for word and sentence alignment and statistical machine translation. A corpus is a collection of written text, or sometimes transcribed speech, stored electronically. It is meant to be representative of the respective language and/or domain it contains, although it can never be completely representative due to the enormous variation inherent in language. It may consist of prose, newspaper material or technical texts; a given corpus often represents a specific genre. Two types of corpora can be distinguished: corpora containing one language (*monolingual corpora*) and corpora containing two or more languages (*multilingual corpora*). A multilingual corpora consisting of two languages is called a *bilingual corpora*. Multilingual corpora can be divided into parallel and non-parallel corpora.

Parallel corpora contain one source document, which is the original text, and one or more target documents in other languages. A non-parallel corpus consists of a single text document, or two or more non-overlapping text documents in the same language. Several bilingual and multilingual corpora have been produced during the

---

<sup>1</sup>Another example of a 'fuzzy' alignment, from the gold standard of the Bellow bitext: 'unrelated - inte tillhör hans släkt' (Tiedemann (2003)

last decade. Examples are the English-French Canadian Hansard corpus, Lancaster's ITU corpus and the MULTEXT-East corpus for eastern European languages. The Hansard corpus contains one target language whereas MULTEXT is a multilingual corpus containing several target languages (Songlin Piao 2001).

## 4.4 Methods for word alignment

The first word alignment prototypes were developed in the early 1990's by Brown et. al (1990) and Gale and Church (1991). Recent development in word alignment includes *GIZA++* (Och 2002); an implementation and refinement of translation models originally proposed by IBM (Brown et. al 1993), and the *Clue aligner* (Tiedemann 2003) developed within the PLUG project at Uppsala University. While *GIZA++* is statistics-based, the *Clue aligner* combines both statistical and linguistic resources. Linguistic resources are for example *declarative clues*, which are clues to associations between word pairs from related word classes.

In experiments conducted with the *Clue aligner*, translation probabilities from *GIZA++* were included, and this combination of statistical and linguistic clues resulted in an impressive performance rate of above 80% in precision and recall <sup>1</sup> (Tiedemann 2003).

Other approaches to word alignment are also being investigated by researchers. Systems such as the *Clue aligner* demand a certain degree of contextual knowledge and are time-consuming, as the corpora involved are often very large.

### 4.4.1 The association approach vs. the estimation approach

The two main types of word alignment are the *association approach* and the *estimation approach*. Both approaches make use of statistics. The *association approach* is built on correspondence measures and originates from early studies on lexical analysis of parallel corpora (Tiedemann 2003). The *estimation approach* is based on probabilistic translation models from statistics and is sometimes called *statistical alignment*. Alignment models used in the estimation approach are estimated from parallel corpora, an approach heavily influenced by statistical machine translation, incorporating the *noisy channel model* (Shannon 1948).

#### **The association approach**

The basic steps in aligning with the association approach are:

*Lexical segmentation*: identification of boundaries of lexical items in source and target language.

*Correspondence*: possible translation relations between lexical items are identified according to correspondence criteria, resulting in an association dictionary with association scores for every entry.

---

<sup>1</sup>for information on precision and recall, see 4.6

*Alignment and extraction*: the most reliable translations from the association dictionary are marked in the alignment, often by using a "best first" search algorithm combined with linguistic and heuristic constraints. The aligned words are then extracted into a bilingual translation dictionary (Tiedemann 2003:13).

Statistical information is applied in association approaches by way of *co-occurrence measures*: testing if two words co-occur more frequently than expected if they co-occur by chance only. An example of this method is the test metric *t-score*, which is derived from the *t-test* in statistics theory. Other co-occurrence measures used in association approaches are the *dice coefficient*, which measures the correlation between two events, and *point-wise mutual information*, which originates from information theory (Tiedemann 2003).

String similarity measures are also applied, resulting in findings of *cognates*. This technique is especially useful when comparing closely related languages. External alignment resources such as machine-readable bilingual lexicons can also be of use as they contain definitions of common relations between words and phrases (Tiedemann 2003).

### The estimation approach

Estimation approaches make use of probabilistic alignment models estimated from parallel corpora. The models are often derived from statistical machine translation, i.e. the translation models introduced by IBM researchers in "*The mathematics of statistical machine translation: Parameter estimation*" (Brown et. al 1993).

Alignment in Statistical Machine Translation (SMT) is modelled as having a sequence of hidden connections, where each word in a target language string **t** is connected to *not more* than one word in the source language **s**. Consequently, this is called the *directional alignment model*. To handle words for which there is no possible equivalent in the other language, an *empty word* is put at position 0 in the source language string (Tiedemann 2003).

#### 4.4.2 The Clue aligner

The automatic word aligner I have used in this thesis is the Clue aligner (Tiedemann 2003), which combines various *association clues*. Inspired by the "greedy" word alignment approach of the UWA (Uppsala Word Aligner) (Tiedemann 2001a), it utilizes both statistical resources and linguistic knowledge, which are regarded as *clues* to relations between words and phrases. The linguistic information, along with contextual features, can point out translational correspondences between words in bitexts<sup>1</sup>. Examples are spelling similarities, word class information, word position, morpho-syntactic features and syntactic features.

Tiedemann (2003) performed linking experiments with the Clue aligner on three parallel corpora from the PLUG project (Sågvall Hein 2002). The experiments were conducted in four steps.

The first series of alignment experiments use *basic clues*: alignment clues derived from the bitext, without any help from training material or external sources. The

---

<sup>1</sup>a bitext is a bilingual parallel corpora

measures used are based on string similarity, co-occurrence and translation probabilities yielded by statistical alignment, i.e. GIZA and GIZA++. Examples of *string similarity measures* applied by the Clue aligner are *dice coefficient* and *LCSR* (Longest Common Sub-sequence Ratio).

The second series adds *declarative clues*; pre-defined clues such as associations between word pairs from related word classes; for example, a definitive article and a noun in an English text are likely to correspond to a definite noun in Swedish. Declarative clues are static and independent of the bitext that is linked.

Sets of declarative clues tested in the Clue aligner are clues from pairs of part-of-speech tags, part-of-speech clues that match multi-word units (MWU)<sup>1</sup>, and a set of chunk label pairs.

In the third series, *dynamic clues* learned from linguistically enriched (for example part-of-speech tagged) bitext segments in pre-aligned corpora are applied. The fourth and last series investigates the effect of different search strategies. The resulting alignments were evaluated automatically.

The evaluation of the experiments reveal that string matching clues yield the least impressive results for word alignment, while the effect of declarative clues is dependent on what set of basic clues they are added to. Dynamic clues may improve precision while the effect on recall is secondary (Tiedemann 2003).

## 4.5 Applications of word alignment

The bilingual data produced by word alignment can be of use in commercial machine translation systems, in the creation of translation support, various multilingual computerized resources such as bi- and multilingual translation lexicons, translation memories (Tiedemann 2003) and terminology databases for NLP systems. As an example, the lexicon of the rule-based machine translation platform MATS (Sågval Hein et. al 2004) was extended by more than 7000 lexemes derived from word alignment results (Tiedemann 2003). Another example is *Wordnet* and its multilingual version *Eurowordnet* (Salkie 1999). Extracted lexicons are mostly built on word alignment results based on association approaches, excluding uncertain relations and grammatical functions, as the identified translated relations are to be used free of context (Tiedemann 2003).

The other main application of word aligned texts is in the field of statistical machine translation, an area which is closely related to and which can benefit from word alignment.

### Statistical Machine Translation (SMT)

Statistical machine translation is based on statistical methods and ideas from information theory developed by Claude Shannon (1948), i.e. the *noisy channel model* (Brown et al. 1990). The noisy channel model can be explained as having the source **s**, which by the time it gets on paper is corrupted/distorted by noise and becomes the target **t**. The idea is to recover the most likely **s**, and this is accomplished by reasoning about what kinds of things are said in the source language and how it is turned into

---

<sup>1</sup>"[...]word sequences and word groups, which express structural and conceptual units such as complex noun phrases, phrasal verbs, idiomatic expressions[...]that should not be split up in the alignment process" (Tiedemann 2003:18)

the target language (Knight 1999). Probabilities applied in the noisy channel model are:

- \*  $P(\mathbf{s})$  - the chance that  $\mathbf{s}$  happens (*a priori probability*)
- \*  $P(\mathbf{s}|\mathbf{t})$  - the chance of  $\mathbf{s}$  given  $\mathbf{t}$  (*conditional probability*)
- \*  $P(\mathbf{s},\mathbf{t})$  - the chance that both  $\mathbf{s}$  and  $\mathbf{t}$  happens (*joint probability*).

If there is no influence between  $\mathbf{t}$  and  $\mathbf{s}$ , the joint probability may be written as  $P(\mathbf{s},\mathbf{t}) = P(\mathbf{s}) * P(\mathbf{t})$  (Knight 1999:1).

When the noisy channel model is applied in machine translation, the source language  $\mathbf{s}$  and the target language  $\mathbf{t}$  represents strings consisting of sentences or words. The language string  $\mathbf{s}$  is transmitted through a noisy channel and transformed into a string  $\mathbf{t}$  in the target language.

The task of the noisy channel model is to find the input string that yielded the output string (Tiedemann 2003). The risk of error is minimized by choosing the string  $\mathbf{s}$  that is the most probable given  $\mathbf{t}$ , in other words, the most likely translation from one language to another is sought. If Swedish is the source language, and English the target language, given a Swedish sentence we seek the English sentence that maximizes the probability  $P(\mathbf{s}|\mathbf{e})$ . Using Bayes's theorem, the probability of the input string  $\mathbf{s}$  given  $\mathbf{e}$  is written as:

$$P(\mathbf{s}|\mathbf{e}) = \frac{P(\mathbf{s}) * P(\mathbf{e}|\mathbf{s})}{P(\mathbf{e})}$$

As  $P(\mathbf{e})$  is independent of  $\mathbf{s}$  and constant for all possible strings  $\mathbf{s}$ , the theorem can be simplified by saying that it suffices to choose the  $\mathbf{s}$  that maximizes the product  $P(\mathbf{s}) * P(\mathbf{e}|\mathbf{s})$ . The *argmax* expression is used in combination Bayes's theorem for formulating the most likely translation (Knight 1991); what has been called the Fundamental Equation of Machine Translation (Brown et al. 1993:265):

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{s}) * P(\mathbf{e}|\mathbf{s})$$

The probability  $P(\mathbf{e}|\mathbf{s})$  is a *translation model*, which is supposed to be estimated from sentence-aligned parallel corpora. However, it is not possible to estimate  $P(\mathbf{e}|\mathbf{s})$  directly from a corpus, as the majority of segments in parallel corpora are unique, and many possible sentences will not appear in the training corpora (Tiedemann 2003). To solve this problem, the translation model is “[...]decomposed into distributions of smaller units, which recur more often in the training data and are more likely to appear again in unseen data (Tiedemann 2003:21).”

Decompositions in statistical machine translation are often based on the five translation models and algorithms introduced by IBM. In short, the idea is to start with a simple model and then progress to more advanced models where the output of the simpler models can be used. Each model computes the conditional probability  $P(\mathbf{f}|\mathbf{e})$  ( $\mathbf{f}$  for the source language French and  $\mathbf{e}$  for the target language English). A large number of parameters are estimated in a process known as *training* and then applied to an algorithm. The first model is a simple word translation model using co-occurrence of corresponding words in sentence aligned bitext segments, while the

remaining models add various dependencies (Brown et al. 1993). One of the most recent applications based on IBM's models is the statistical alignment tool GIZA, which was developed in 1999 at the Johns-Hopkins University and updated in a revised version as GIZA++ (Och 2002).

## 4.6 Evaluation metrics

The most common metrics for measuring the performance of NLP systems such as word aligners are precision and recall (Merkel et. al 1999b). In evaluation of word alignment, precision represents the number of correctly aligned items in proportion to the number of aligned items. Recall represents the number of correct results compared to the number of correct items in total. In Tiedemann (2003), an *F*-measure is also presented, which combines measures for both precision and recall in order to capture the overall performance.

$$P = \frac{|aligned \cap correct|}{|aligned|}, R = \frac{|aligned \cap correct|}{|correct|}, F = \frac{2 * P * R}{P + R}$$

Figure 4.1: Evaluation metrics (Tiedemann 2003:26)

## 4.7 Lemmatization

### 4.7.1 Definition of a lemma

Allén (1971) defines a lemma as a group of word forms within a word class which are derived from the same paradigm and have basically the same pronunciation. Semantic differences are not taken into account. To automatically derive the lemma from a word form, the word form is typically morphologically analyzed. The result should be the base, dictionary form of the word, together with morphosyntactic information given by affixes. Usually only content words; verbs, nouns, adjectives and participles, are lemmatized.

Some NLP systems define the word stem as a lemma marker (Karlsson 1990), while this thesis follows the guidelines of the Uppsala Chart Processor (UCP) (Sågval Hejn 1982, 1997), which differentiates between lemma and word stem in order to separate lemmas that have identical stems. Generally, the more inflective a language is, the harder it is to formulate and evaluate rules for lemmatization (Paskaleva 2003). Less inflective languages, such as English, are thus easier to process.

### 4.7.2 NLP systems for morphological analysis

Lemmatization is part of morphological analysis, which forms the basis for many applications in NLP systems, such as syntax parsing, machine translation and automatic indexing (Lezius et al. 1998). Morphological analysis is a useful resource for NLP

systems, as the systems can store information about a word that is keyed by lemmas instead of full word forms. Also, the lemmas of words can be used to generalize across the corresponding word forms (Minnen et al. 2001). Several systems for morphological analysis have been created for several languages, although the lexicons for most systems are quite small, and many systems are not freely available (Lezius et al. 1998).

An important system for morphological analysis involving Swedish was constructed by Karlsson (1990); a constraint grammar parser which includes a morphological analysis step for Swedish, English and Finnish, designed according to Koskenniemi's (1983) two-level morphology model.

The two-level morphology model describes phonological alterations in finite-state terms. Each rule has a constraint for correspondence between the surface and lexical level of words, and states the environment in which the correspondence is allowed. For example, one rule might describe the *y* -> *ie* alteration in English plural nouns (Karttunen et. al 2001).

For the Swedish language there are a few NLP systems that generate lemmas. An example is the Uppsala Chart Processor (Sågval Hein 1982, 1997), which is a part of the MATS (Sågval Hein et. al 2004) system. The built-in morphological analysis in UCP generates lemmas from pattern-words. Another example is the lemmatizer developed by Kokkinakis et al. (1997) within the AVENTINUS project at Gothenburg University, currently available in an on-line interface which unfortunately only has default input files. It yields lemmas for content words given an input of word forms only and seems to work fairly well, although the system has problems differentiating between word classes in some cases.

For the English language, there is the *Morph* system (Carroll et al. 2001). It is freely available and includes tools for morphological analysis and generation, incorporating *Morpha* (Carroll et al. 2001a), a morphological analyzer for the English language that returns the lemmas given an input consisting of the word form and part-of-speech, i.e. the input has to be tagged for word form.

At present, *Morpha* only returns lemmas for nouns and verbs, not adjectives or participles. For German, there is a similarly titled morphological analyzer: *Morphy*, which includes a context-sensitive lemmatizer (Lezius et al. 1998).

## 5 Method

Recent developments in word alignment, for example the *Clue aligner* and *GIZA++* tools, have set high performance standards. In this thesis, the possibility of improving precision and recall measures by using lemmas instead of word forms as input to word alignment is examined. Word forms within the same paradigm can be generalized into lemmas, which may result in a higher percentage of correctly linked and recognized items.

I chose to extract lemmas from existing lexical resources, from which selected information is stored in new lexicon files. A script created in Perl then mines the new lexicon files for lemmas to replace word forms with, while the remaining Swedish word forms for which no lemmas are found are run through the UCP compound analyzer (Sågvall Hein 1982, 1997). The result is aligned by the Clue aligner and evaluated automatically. The evaluation is contrasted to a *word form* alignment evaluation.

Furthermore, the possibilities of extracting the produced alignments for use in lexicons are examined, more specifically as a lexical resource addition to the MATS (Sågvall Hein et. al 2004) system. The result is two translation dictionaries consisting of lemma entries.

The texts used originate from experiments conducted with the Clue aligner by Tiedemann (2003) which were evaluated in precision and recall using *gold standards*<sup>1</sup>. The texts consist of four PLUG-XML documents of two parallel corpora from the PLUG (Sågvall Hein 2002) project.

### 5.1 Deriving lemma forms

#### 5.1.1 Deriving Swedish lemma forms

No reliable NLP system exists for the extraction of lemmas of Swedish words. Thus, I have mined available lexical resources obtained from the Department of Linguistics and Philology at Uppsala University for lemma forms.

Lexicon files are created for use in a Perl script with selected information from the resources. The resulting files contain three columns separated by tabs; the first contains the word form, the second the corresponding lemma with extension, and the third a SUC (Stockholm Umeå Corpus) (Ejerhed et. al 1997) code, all sorted in lower case and including only single occurrences of every word form.

The lexical resources are the following:

---

<sup>1</sup>a gold standard is a reference alignment

- Text material regarding education information at Uppsala University, collected and processed in the Project Course Database (Pettersson 2005). Every entry in the lexicon has morphosyntactic information divided into lemma, lexeme, technical stem, paradigm and word class (Pettersson 2005). Lemmas with extension, corresponding word forms and grammatical codes are selected and saved in a file.

- The lexicon from the Scarrie (Scandinavian Proof-reading Tools) (Olsson 1999) database. Scarrie contains text material from the Swedish newspapers Uppsala Nya Tidning and Svenska Dagbladet (Olsson 1999). Lemmas with extension, corresponding word forms and grammatical codes are selected and saved in a second file.

- The lexicon from the Swedish translation prototype created for Systran (Gustavii et al. 2003), a machine translation system used by the EU commission (Gustavii et al. 2003). Lemmas with extension, corresponding word forms and grammatical codes are selected and saved in a third file.

- The lexicon from the car domain database developed within the MATS project (Sågvall Hein et. al 2004). Lemmas with extension, corresponding word forms and grammatical codes are selected and saved in a fourth and last file.

An example of a Swedish word form-lemma lexicon file:

**Table 5.1:** Lexicon file example

<b>Word form</b>	<b>Lemma</b>	<b>Morphosyntactic code</b>
adoptionsfrågor	adoption+fråga.NN	NNUPIG
adoptionen	adoption.NN	NNUSDB
adoptionens	adoption.NN	NNUSDG
adoptionerna	adoption.NN	NNUPDB
adoptionernas	adoption.NN	NNUPDG

### 5.1.2 Deriving English lemma forms

Lemmatizing English word forms is less complex than lemmatizing Swedish word forms, as English is a less inflective language. The *Morph* (Carroll et. al 2001a) system for morphological analysis can be applied for automatic lemmatization. The only drawback is that Morph does not analyze adjectives or adverbs; only nouns and verbs.

The English corpora are made into single-line format, with each line containing a word form and its grammatical code, and then processed by Morpha. The resources for English lemmas are the following:

- Lemmas from Morpha tagged with the *TnT*-tagger (Brants 2000) are saved together with the the lemma extensions and word forms from the English corpora in respective lexicon file.

- Lemmas are extracted from the Systran lexicon (Gustavii et al. 2003), Project Course Database (Pettersson 2005) and the car domain database developed within

the MATS project (Sågvall Hein et. al 2004) and saved in respective lexicon file.

- If there remains any English adjectives for which no lemmas are found, they are lemmatized by hand.

## 5.2 Construction of Perl script

To derive lemmas from the newly created lexicon files, a script has been implemented in the programming language Perl. The files containing word forms, corresponding lemmas, grammatical codes and extensions are used as input to the program, along with the word form PLUG-XML documents.

An effort to keep the lemmatizing process as automatic as possible is made; manual corrections of tokenization or tagging errors in the resulting or the original XML documents are avoided. Also, the sentence id's in the XML files are saved as they are for later use; only word forms are substituted.

The Perl program is updated throughout my work, as new problems to be dealt with surface from time to time. Roughly, the following steps are performed in the construction of the Perl script:

- 1) Each lexicon file is read. Word forms, lemmas, and for the Swedish bitexts also morphosyntactic SUC (Ejerhed et. al 1997) codes, are saved in variables.

- 2) The corpus file to be lemmatized (in XML format) is read. Each word form and each corresponding grammatical code are saved in variables. Other parts of each line is saved in separate variables. All the word forms are translated to lower case so that they can be mapped to the word forms in the lexicon files. Words that include upper case letters are saved in separate variables and restored to their original configuration after comparison with the lexicon files, in order to remain as close as possible to the word form XML-documents.

- 3) Word forms in the XML file are mapped to word forms in the lexicon files. If a lemma is found for the word form, it replaces it in the XML file. If the lemma is not found in the first lexicon file, the program moves on to the next file. If an identical lemma is found for a word form which has already been given its lemma, it will be ignored.

- 4) Entries which have been given more than one lemma are saved and processed separately in a subroutine. The grammatical code of each word form is compared to the grammatical codes or extensions of the lemmas it has been connected to. The correct lemma is selected.

- 5) The result is printed to file. It will correspond exactly to the original XML file, but with all word forms replaced by lemmas.

- 6) Words for which no lemmas are found are saved in an "error" file for further analyzing.

## 5.3 Compound analyzing

The remaining Swedish word forms for which no lemmas have been found are likely to include a great number of compounds. Especially the Scania corpus contains many compounds of technical nature. Therefore, they are run through the UCP (Sågwall Hein 1982, 1997) compound analyzer. The resulting entries are run through a script (Åberg 2003) that for each compound chooses the best analysis given by the UCP analyzer. Lemmas in compounds are separated by a plus: *ventil+hylsa.NN*.

## 5.4 Aligning with the *Clue aligner*

The lemmatized PLUG-XML documents are used as input to the Clue aligner. Each corpus is represented by a source and target file, in the case of the Bellow corpus in English-Swedish and in the case of the Scania corpus in Swedish-English. The documents are incorporated into an existing sentence alignment (Tiedemann 2003), as there is no reason to make a new sentence alignment with lemma forms. The sentence alignment is then processed by the word aligner.

The various *clues* mentioned earlier are applied automatically in the linking process, and the output result consists of links where each link contains a *lexPair* (Tiedemann 2003), i.e. the source language word and the translated word along with a decimal number representing link certainty. New word form alignments are also made, as the Clue aligner has been updated since the initial experiments reported in Tiedemann (2003).

## 5.5 Evaluation of word links

Evaluation of the word links produced will be performed automatically using *gold standards*; reference alignments created manually by aligning chosen segments from a bitext. While it can be difficult to create a truly representative gold standard, once created it can be re-used many times (Tiedemann 2003).

After running the Clue aligner, results will be evaluated in precision, recall and *F* measures by using the gold standards from the PLUG (Sågwall Hein 2002) project. Produced using a word sampling method and the PLUG link annotator (Merkel et al. 1999a), five hundred words were chosen from different genres for Swedish-English and Swedish-German. The standards include 'fuzzy' links, 'null' links<sup>1</sup> and multi-word unit links. However, as the PLUG gold standards consist of word forms, they will have to be lemmatized in order to be applicable for lemma alignments.

The resulting measures from the evaluation of the linked lemmatized corpora will be compared against the newly linked versions of the word form XML documents.

## 5.6 Application of extracted data

Data from alignments will be used to create a bi-lingual lexicon for possible use in the MATS (Sågwall Hein et. al 2004) system as a lexical resource addition. MATS is a rule-based machine translation system developed at the Department of Linguistics

---

<sup>1</sup> words which are not translated

and Philology at Uppsala University, based on the research prototype of the transfer-based system MULTRA and intended for industrial use (Sågvall Hein et al. 2004).

To be able to include an extension for every lemma in the resulting translation dictionaries, extensions are to be included in separate aligning experiments. Lexicons are extracted from word alignments using an UPLUG (Sågvall Hein et. al 2002) tool, *dic-freq*. The dictionary gives frequencies for each translation pair, i.e. how many times it has been linked, starting with the pairs that have only been linked once and ending with the most common word pairs.

## 6 Pre-processing and lemmatizing

### 6.1 Corpora background

The bitexts used as input for the Clue aligner are the literary text "*To Jerusalem and back: a personal account*" (a novel by Saul Bellow) in English-Swedish, and the technical text Scania95 in Swedish-English from PLUG (Sågvall Hein 2002). PLUG is a co-operative project between the Department of Computer and Information Science at Linköping University, the Department of Linguistics and Philology at Uppsala University, and the Department of Swedish Language at Gothenburg University. The project focuses on the generation of translation data from sentence aligned bitext, with Swedish as the source or target language. A corpus of four languages: Swedish, English, German and Italian was yielded. Three genres are represented; technical text, political text, and literary text.

The Bellow novel was supplied by Linköping University and originates from the Swedish Language Bank<sup>1</sup> in Gothenburg, while the Scania corpus was supplied by Uppsala University. The Scania corpus contains texts provided by Scania CV AB, originally for another study which aimed to establish a controlled vocabulary for truck and bus maintenance (Sågvall Hein 2002).

### 6.2 Pre-processing and format of corpora

As it was decided to lemmatize the corpora *after* tokenization and part-of-speech tagging, instead of tagging raw text files and then lemmatizing, the PLUG-XML documents of the two corpora were kept in their original format. Tokenization, tagging and conversion into XML-format had already been performed by Tiedemann (2003). There are errors to be found in the tagging, as all mark-up has been made automatically. However, the original tagging could not be improved automatically; an attempt to re-tag the English texts with the *TnT*-tagger was made, but the results still contained a number of errors. This led to my decision of using the PLUG-XML documents as input instead of going back to the raw texts.

---

<sup>1</sup>see <http://spraakbanken.gu.se/>

**Table 6.1:** Corpora

<b>Corpora</b>	<b>Language pair</b>	<b>Number of words</b>
Bellow	English → Swedish	132066
Scania95	Swedish → English	385289

The English texts were tagged and parsed with the NLP tool *Grok*<sup>1</sup> and the Swedish texts were tagged with the *TnT-tagger* and parsed with Beáta Megyesi's (Megyesi 2002) part-of-speech parser (Tiedemann 2003).

An example from a PLUG-XML file:

```
<s id="1">
  <chunk id="c-1" type="NP">
    <w span="0:8" pos="NN" id="w1.1">Security</w>
    <w span="9:8" pos="NNS" id="w1.2">measures</w>
  </chunk>
  <chunk id="c-2" type="VP">
    <w span="18:3" pos="VB" id="w1.3">are</w>
  </chunk>
  <chunk id="c-3" type="ADJP">
    <w span="22:6" pos="JJ" id="w1.4">strict</w>
  </chunk>
  <chunk id="c-4" type="PP">
    <w span="29:2" pos="JJ" id="w1.5">on</w>
  </chunk>
  <chunk id="c-5" type="NP">
    <w span="32:7" pos="NNS" id="w1.6">flights</w>
  </chunk>
  <chunk id="c-6" type="PP">
    <w span="40:2" pos="TO" id="w1.7">to</w>
  </chunk>
  <chunk id="c-7" type="NP">
    <w span="43:6" pos="NNP" id="w1.8">Israel</w>
  </chunk>
</s>
```

Files encoded in the PLUG XML (Sågvalld Hein 2002) format start off with a short header followed by a body of sentence-aligned *sub-corpora*. Each sentence is marked with a *sentence id* and sentences are divided into phrases containing the words they are made up of and corresponding grammatical labels. The words in the Swedish corpora have morphosyntactic SUC (Ejerhed et. al 1997) codes.

## 6.3 Extraction of lemmas

The extracted word forms and lemmas from the lexical resources mentioned resulted in five lexicon files for Swedish containing material from the Project Course Database (Pettersson 2005), the Scarrie (Olsson 1999) database, an agricultural lexicon developed with texts from Systran (Gustavii et al. 2003), the output from the UCP compound analyzer (Sågvalld Hein 1982, 1997), and the car domain database (Sågvalld Hein et. al 2004). After being sorted to include only unique entries, the files contain the following number of word forms with respective lemmas:

---

<sup>1</sup>Available at <http://grok.sourceforge.net/>

**Table 6.2:** Lexicon files for Swedish

Source	Amount
Scarrie	784897
Project Course Database	74993
Systran	40936
Output from compound analyzer	4140
Car domain database	125333

The English lemmas yielded by the *Morph* (Carroll et al. 2001a) analyzer were tagged with the *TnT* (Brants 2000) tagger using the Wall Street Journal (WSJ) model (Brants 2000) and sorted uniquely, one file for each corpora, together with corresponding word forms. All adverbs, adjectives and participles were removed, as they were not analyzed correctly by TnT.

English lemmas were also extracted from the agricultural lexicon (Gustavii et al. 2003), the Project course database (Pettersson 2005) and the car domain database (Sågwall Hein et. al 2004). Additionally, all words tagged in the English PLUG-XML documents as adverbs, adjectives and participles which were not given lemmas by the Perl script were lemmatized by hand. After being sorted to include only unique entries, the English lexicon files contained the following number of word forms entries with lemmas:

**Table 6.3:** Lexicon files for English

Source	Amount
Morpha output 1 (Bellow corpus)	8114
Morpha output2 (Scania corpus)	5880
Systran	7932
Project Course Database	16837
Adjectives, adverbs, participles lemmatized manually	61
Car domain database	28603

As many word forms and lemmas can be found in more than one file, there is a certain degree of overlapping. However, this was taken into account when constructing the Perl script and constitutes no problem.

## 6.4 Replacing word forms with lemmas

Swedish and English word forms in the PLUG-XML corpora were replaced by lemmas automatically by running the newly constructed Perl script. Lemmas were found for most of the word forms. The basis for the following results were produced by dividing the number of tokens lemmatized with the number of tokens in plain text versions of the XML source files. A comparison yields the following figures:

**Table 6.4:** Success rate for lemmatized corpora

Corpus	Language	Lemmatized word forms
Bellow	English	98%
Bellow	Swedish	86 %
Scania95	English	88%
Scania95	Swedish	84%

As the results show, the lemmatizing was more successful on the English texts than on the Swedish texts. One reason for this may be that English is less inflective than Swedish and thus easier to lemmatize. The Bellow bitext also shows better results than the Scania95 bitext, probably because the Scania95 corpus includes many more symbols and non-alphabetic characters than the Bellow corpus, as it is a technical manual.

In many cases where lemmas are not found it is the result of incorrect markup of the original text files, where non-alphabetic characters have mistakenly been attached to word forms. In other cases, the word forms are very uncommon words with a strong contextuality, compounds, names of places and persons or cardinal numbers. Also, the Bellow corpus contains words in foreign languages while the Scania corpus contains many errors in spelling which I choose to leave out of the lexicon files.

Tagging errors are present in the XML documents. Also, sometimes my lexicon files contained different lemmas for the exact same word form with the same tag. These issues were considered when constructing the Perl script, but to make sure that each word form was given only one lemma and not more or less, the output from the Perl script was pasted into a file with the corresponding original XML file, so that errors could be detected and corrected manually by removing error lemmas from the lexicon files.

Swedish words for which no lemmas were found were run through the lemmatizer from Gothenburg University (Kokkinakis et. al 1997). As the output contained some errors, the results were reviewed and manually corrected where needed. Nothing further was attempted with the remaining English word forms, as most of them were foreign words, symbols or incorrectly tokenized.

To measure lexical variation in each lemmatized bitext, type-token ratio measures were calculated, the types in this case consisting of lemmas. The results were yielded by dividing the unique lemmas with the total number of tokens for each file. Clearly, there is greater lexical variation in the Bellow bitext:

**Table 6.5:** Type-token ratio for lemmatized corpora

<b>Bellow corpora</b>	
<i>English</i>	8566/77582 = 0,11
<i>Swedish</i>	9612/73747 = 0,13
<i>Total</i>	18178/151329 = 0,12
<b>Scania95 corpora</b>	
<i>Swedish</i>	14509/188361 = 0,077
<i>English</i>	8634/239155 = 0,036
<i>Total</i>	23143/427516 = 0,054

## 7 Results and evaluation of word links

The lemmatized PLUG-XML documents were inserted into the existing sentence alignments made by Tiedemann (2003), which I converted from an older format to PLUG-XML format. Word alignment was then performed on the sentence alignments of the lemmatized Bellow bitext and the Scania95 bitext. After this, sentence alignments with word forms were aligned. The most recent, updated version of the Clue aligner (Tiedemann 2003) was used <sup>1</sup> in both cases. The Clue aligner gives the opportunity of including a lexicon as reference when aligning, but no lexicons were included in my experiments.

In order to evaluate the result of the lemma links, the gold standards from the PLUG (Sågvald Hein 2002) project were lemmatized by hand to assure correctness, with the lexicon files as reference. The majority of the links in the gold standards are regular, but 'fuzzy' links and null links are also included. A PLUG evaluation tool was used to compare the corpora against the gold standards.

The results consist of correct links, partially correct links, incorrect links and average scores for each type of link. Missing links, in cases where the aligner has failed to find any link at all, are also reported.

A partially correct link includes at least one correct word on both sides of a link. The partiality, which is a common phenomenon in word alignment, results from the different possibilities of multi word units (Tiedemann 2003). As partial links cannot be evaluated in standard precision and recall, a special measure, *PWA*, is included.

Precision and recall measures is presented for regular and 'fuzzy' links along with *PWA* measures for partial links. An overall measure, the *F*-measure, is also presented, which combines measures for both precision and recall. The measures represent the overall performance of the Clue aligner, i.e. the combined results of the various clues applied in the linking process.

**Table 7.1:** Linking results for parallel Bellow corpora

	<b>Lemmatized corpora</b>
<i>recall</i>	86.65% (regular: 89.50%,fuzzy: 39.88%,pwa: 80.76%)
<i>precision</i>	80.80% (regular: 83.62%,fuzzy: 34.50%,pwa: 80.76%)
<i>F</i>	83.62% (regular: 86.46%,fuzzy: 36.99%,pwa: 80.76%)
	<b>Word form corpora</b>
<i>recall</i>	86.47% (regular: 89.47%,fuzzy: 37.28%,pwa: 80.93%)
<i>precision</i>	80.99% (regular: 83.95%,fuzzy: 32.52%,pwa: 80.93%)
<i>F</i>	83.64% (regular: 86.62%,fuzzy: 34.74%,pwa: 80.93%)

<sup>1</sup> Available at <http://sourceforge.net/projects/uplug>

**Table 7.2:** Linking results for parallel Scania95 corpora

<b>Lemmatized corpora</b>	
<i>recall</i>	79.37% (regular: 80.96%,fuzzy: 54.29%,pwa: 71.76%)
<i>precision</i>	74.41% (regular: 76.07%,fuzzy: 48.51%,pwa: 72.55%)
<i>F</i>	76.81% (regular: 78.44%,fuzzy: 51.24%,pwa: 72.15%)
<b>Word form corpora</b>	
<i>recall</i>	77.53% (regular: 78.56%,fuzzy: 61.32%,pwa: 70.65%)
<i>precision</i>	73.91% (regular: 75.30%,fuzzy: 52.45%,pwa: 71.82%)
<i>F</i>	75.68% (regular: 76.89%,fuzzy: 56.54%,pwa: 71.23%)

A comparison of the evaluation of links yielded from lemmatized corpora and links yielded from word form corpora verifies that word linking results may be improved by using lemmatized corpora. The evaluation of the lemma alignments shows improvement in both precision and recall for the Scania bitext and improvement in recall for the Bellow bitext.

An earlier linking experiment yielded a 75.76% precision measure for the Scania bitext, but the recall was much lower. Similarly, an earlier alignment of the Bellow bitext resulted in a marginally better precision result, 80.87% but lower recall. Generally, high recall measures result in lower precision measures, and high precision measures result in lower recall measures (Tiedemann 2003).

A benefit of linking lemma forms instead of word forms, as shown in the recall measures, is that more items are found. The lemma alignment for the Scania bitext yielded an increase in partially correct links, while measures for 'fuzzy' links were improved for the Bellow bitext.

Precision decreased by 0,19% for the Bellow lemma links compared to the word form links. Other than this, the results were the best for lemmatized corpora. The increase in recall is less than dramatic, but still substantial considering the high level of success already achieved using word form corpora.

In the case of the Scania bitext, lemmatization improved the recall measure by 1,84%, the precision measure by 0,5%, and the *F* measure yields an overall improvement of 1.13%. Recall for the Bellow bitext was improved by 0.18%.

## 8 Application of extracted data

### 8.1 Extraction of lemma dictionaries

Translation dictionaries were extracted from links in the lemmatized corpora alignments. As a dictionary require lemma extensions for each word pair, new versions of the Bellow and Scania95 PLUG-XML files including extensions were produced. Each word form in the original XML files (Tiedemann 2003) was replaced by a lemma with corresponding extension. To make sure that the lemmas would get basically the same extensions in both Swedish and English, the set of extensions used in MATS (Sågvall Hein et. al 2004) were applied on both the Swedish and English corpora. As the English extensions are in the Penn Treebank<sup>1</sup> tagset, they had to be converted to match the MATS format. The English tagset is more extensive than the MATS tagset; consequently some generalizations had to be made. For example, the Penn Treebank tag set has an extension for participles which does not exist in the Swedish tag set. Also, the *Tnt* (Brants 2000) tagger had incorrectly marked many words beginning with capital letters, in some cases the first word of sentences, as proper nouns. This problem was attended to as much as possible.

```
sventscan5    säkerhet.NN    safety.NN
sventscan5    service+verkstad.NN    service.NN workshop.NN
sventscan7    varning+signal.NN    warning.NN signal.NN
sventscan9    mekaniker.NN    mechanic.NN
sventscan28   skada.NN        injury.NN
sventscan28   vara.VB be.VB
sventscan28   med.PP with.PP/SN
sventscan28   verkstadsarbete.NN    workshop.NN activity.NN
sventscan28   att.IE to.IE
sventscan141  eller.CN        or.CN
sventscan741  draga.VB        tighten.VB
sventscan2110 felkod.NN       fault.NN code.NN
sventscan3014 skada.NN        damage.VB
sventscan3317 motor.NN        the.AL engine.NN
sventscan3317 isbildning.NN  ice.NN
sventscan3520 enskild.AV      individual.AV
sventscan3754 kolv.NN piston.NN
sventscan3807 vev+axel.NN    crankshaft.NN
sventscan4029 justera.VB      adjust.VB
```

**Figure 8.1:** Extract from Scania lemma dictionary in Swedish-English

<sup>1</sup>available at <http://www.cis.upenn.edu/treebank/home.html>

For reference, here is an extract of a dictionary based on the Scania word form parallel corpora.

```

sventscan5      Sakerhet      Safety
sventscan5      serviceverkstaden  service workshops
sventscan7      Varningssignaler  Warning signals
sventscan9      mekaniker      mechanics
sventscan28     skada  injury
sventscan28     är  is
sventscan28     med  with
sventscan28     verkstadsarbete  workshop activity
sventscan28     att  to
sventscan141    eller  or
sventscan741    dras åt be tightened
sventscan2110   felkoder      fault codes
sventscan3014   tar skada      damaged
sventscan3317   Motorn The engine
sventscan3317   isbildning    ice forms
sventscan3520   enskild area
sventscan3754   Kolvar Pistons
sventscan3807   vevaxeln      crankshaft
sventscan4029   justera adjust

```

**Figure 8.2:** Extract from Scania word form dictionary in Swedish-English

The resulting lemma dictionary is not a finished dictionary and is likely to include some errors. For example, the Perl script does not differentiate between identical lemmas of different part-of speech in all cases: (*screen.NN* vs. *screen.VB*). Also, as mentioned earlier, the markup of the PLUG-XML files is not fail safe, which in some cases results in incorrect extensions.

It was not known if including lemma extensions would create problems when aligning with the Clue aligner. Evaluation of the resulting word links using a gold standard was attempted, but the UPLUG (Sågvall Hein et. al 2002) evaluation tool did not accept alignments containing lemmas with extension as input. This problem was solved by removing the extensions and then evaluating the link results.

**Table 8.1:** Linking results for parallel Bellow corpora with lemma extensions

	<b>Lemmatized corpora with extensions</b>
<i>recall</i>	86.26% (regular: 89.23%,fuzzy: 37.53%,pwa: 79.90%)
<i>precision</i>	79.94% (regular: 82.97%,fuzzy: 30.21%,pwa: 79.90%)
<i>F</i>	82.98% (regular: 85.99%,fuzzy: 33.48%,pwa: 79.90%)

**Table 8.2:** Linking results for parallel Scania95 corpora with lemma extensions

	<b>Lemmatized corpora with extensions</b>
<i>recall</i>	80.35% (regular: 81.34%,fuzzy: 64.78%,pwa: 71.89%)
<i>precision</i>	74.06% (regular: 75.16%,fuzzy: 56.88%,pwa: 72.88%)
<i>F</i>	77.07% (regular: 78.13%,fuzzy: 60.57%,pwa: 72.38%)

There were no missing links reported at all for the Bellow alignment, while only 5

were missing in the Scania alignment. Apparently the inclusion of lemma extensions had little negative effect on recall, which decreased by only 0.39% for the Bellow bitext and actually increased by 0.98% for the Scania bitext, making 80.35% the best recall result for the latter bitext. Precision was slightly negatively affected in both cases, but not enough for one to be able to draw any safe conclusions. It may be that the inclusion of extensions ruled out the linking of some homographs.

## 9 Discussion

Alignment experiments were conducted featuring lemmatized corpora, and yielded superior results in recall, and in one case precision, compared to word form alignments of the same text material. This indicates that lemma form alignment is a strategy that may result in more recognized words in general and in some cases more correct word links.

However, when comparing the link results of the lemmatized Bellow and Scania95 bitexts, the Scania bitext yielded more improvement. The word form alignment of the Bellow bitext is already very good, with an *F*-value of 83.64% compared to 75.68% for the Scania bitext. Tiedemann (2003) also achieved a better result for Bellow bitext compared to the Scania bitext.

An explanation for the diverging results is that the terminology in the Scania bitext is more consistent than in the Bellow bitext. The type-token ratios (see 6.5) reveal that the Bellow bitext has the highest degree of lexical variation. Many of the words in the Scania bitext recur frequently, resulting in fewer token pairs that are only linked once or twice. Taking these facts into account, it is possible that alignment of lemmatized corpora yields the best results when the corpora does not exhibit great lexical variation.

It should be mentioned that one of the lexical resources used for lemmatization comes from Scania, thus containing the same type of vocabulary as the Scania bitext.

If time and resources had permitted, it would have been useful to lemmatize and link another technical corpus and another literary corpus. Perhaps this might have yielded a similar divergence in results.

A major advantage of aligning lemmatized corpora is that lemma lexicons with part-of speech can be extracted. The inclusion of extensions in linking experiments did not have any significant negative effect on the link results; precision was slightly worse than without extensions, but for one bitext, recall was improved. There are some errors in the lexicons, but the ability to produce a lemma lexicon directly (although error-checking is needed), including part-of speech, out of word linking results saves time and energy; the usual technique for extracting lemma dictionaries is to (often manually) lemmatize the link results after aligning. Word link results from lemmatized corpora may thus be used for simplifying the creation of bilingual lemma form lexicons. What one could ask for is a method to ensure that only correct lemma extensions are delivered, and a method to evaluate the correctness.

# Bibliography

Allén, Sture (1971). *Nusvensk frekvensordbok 1*. Stockholm: Almqvist & Wiksell.

Brants, Thorsten (2000). TnT - A statistical part-of-speech tagger. [Electric]. Accessed: <http://www.coli.uni-saarland.de/thorsten/tnt/> [2005-02-12].

Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Frederick, Lafferty, John D., Mercer, Robert L., Roossin, Paul S. (1990). A statistical approach to machine translation. *Computational Linguistics*, vol. 16:2, p. 79-85.

Brown, Peter F., Mercer, Robert L., Della Pietra, Stephen A., Della Pietra, Vincent J. (1993). The Mathematics of Statistical Machine Translation: Parameter estimation. *Computational Linguistics*, vol. 19:2, p. 263-277.

Carroll, John, Minnen, Guido, Pearce, Darren (2001a). Robust, applied morphological generation. *Proceedings of the 1st International Natural Language Generation Conference*, p. 201-208. Israel: Mitzpe Ramon.

Carroll, John, Minnen, Guido, Pearce, Darren (2001b). Applied morphological processing of English. *Natural Language Engineering*, vol. 7:3, p. 207-223.

Ejerhed, Eva, Källgren, Gunnar (1997). Stockholm Umeå Corpus version 1.0, SUC 1.0. Department of Linguistics, Umeå University.

Gustavii, Ebba, Pettersson, Eva (2003). *Utveckling av ett svensk-engelskt lexikon för maskinöversättning inom jordbruksdomänen*. Uppsala, Department of Linguistics and Philology. (Work rapport/Uppsala University).

Karlsson, Fred (1990). Constraint Grammar as a Framework for Parsing Running Text. In Karlgren, Hans (ed.), *Papers presented to the 13th International Conference on Computational Linguistics*, Vol. 3. Helsinki, p. 168-173.

Karttunen, Lauri, Beesley, Kenneth R. (2001). A short history of two-level morphology. [Electric]. Accessed: <http://www.ling.helsinki.fi/koskenni/esslli-2001-karttunen/> [2005-05-03].

Knight, Kevin (1999). A statistical MT Tutorial Workbook, Prepared in conjunction with the JHU summer workshop. [Electric]. Accessed: [www.isi.edu/natural-language/mt/wkbbk.rtf](http://www.isi.edu/natural-language/mt/wkbbk.rtf) [2005-03-31].

Kokkinakis Dimitrios, Johansson Kokkinakis Sofie (1997). A Robust and Modularized Lemmatizer/Tagger for Swedish Based on Large Lexical Resources. In *Research Reports from the Department of Swedish*, GU-ISS-97-1. Gothenburg University.

Lezius, Wolfgang, Rapp, Reinhard, Wettler, Manfred (1998). A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. *Proceedings of the COLING-ACL 1998*, p. 743-48. San Francisco: ACL/Morgan Kaufmann Publishers

Megyesi, Beata (2002). *Data-Driven Syntactic Analysis - Methods and Applications for Swedish*. Diss., Department of Speech, Music and Hearing, KTH. Stockholm.

Merkel, Magnus, Andersson, Mikael, Ahrenberg, Lars (1999a). The PLUG Link Annotator - Interactive Construction of Data from Parallel Corpora. In L. Borin (ed.) *Parallel Corpora, Parallel Worlds*, Proceedings of Parallel Corpus Symposium, Uppsala, April 22-23, 1999. Amsterdam: Rodopi.

Merkel, Magnus, Ahrenberg, Lars (1999b). *Evaluating Word Alignment Systems*. Linköping, Department of computer and information science. (Work rapport/Linköping University).

Och, Franz Joseph 2002. Franz Joseph Och (homepage). [Electric]. Accessed: <http://www6.informatik.rwth-aachen.de/Colleagues/och/> [2005-04-01].

Olsson, Leif-Jöran (1999). Svenska Scarrie. [Electric]. Accessed: [http://stp.ling.uu.se/jo/scarrie-pub/scarrie\\_sv.html](http://stp.ling.uu.se/jo/scarrie-pub/scarrie_sv.html) [2005-04-04].

Paskaleva, Elena (2003). Compilation and validation of morphological resources. *Balkan conference on informatics*. [Electric]. Accessed: [http://iit.demokritos.gr/skel/bci03\\_workshop/papers/SESSION1-20\\_Paskaleva.pdf](http://iit.demokritos.gr/skel/bci03_workshop/papers/SESSION1-20_Paskaleva.pdf). [2005-04-05].

Pettersson, Eva (2005). *Pilotstudie om maskinöversättning inom ramen för Projekt Kursdatabas - Utveckling av språkliga resurser för ett vetenskapsområde samt utvärdering*. Uppsala, Department of Linguistics and Philology. (Work rapport/Uppsala University).

Salkie, Raphael (2002). How can linguists profit from parallel corpora? In Lars Borin (ed.) *Parallel corpora, parallel worlds*. Amsterdam: Rodopi.

Songlin Piao, Scott (2001). Parallel corpora and alignment. [Electric]. Accessed: <http://www.lancs.ac.uk/staff/piaosl/research/alignment/alignment.htm> [2005-03-31].

Sågvall Hein, Anna (1982). An experimental parser. In *Proceedings of COLING '82*, p. 121-126. Prague: ACADEMIA.

Sågvall Hein, Anna (1997). De morfologiska beskrivningarna i Svenska UCP. Uppsala, Department of Linguistics and Philology.

Sågvall Hein, Anna et. al (1999). Om maskinöversättning. NUTEKs slutrapport av regeringsuppdrag N1999/7189/ITFOU 1999-0922.(Work rapport).

Sågvall Hein, Anna (2002). The PLUG project: parallel corpora in Linköping, Uppsala, Göteborg: aims and achievements. In Lars Borin (ed.) *Parallel corpora, parallel worlds*. Amsterdam: Rodopi.

Sågvall Hein, Anna, Forsbom, Eva, Gustavii, Ebba, Pettersson, Eva, Tiedemann, Jörg, Weinitz, Per (2004). The machine translation system MATS - past, present & future. In *RASMAT'04 (Recent Advances in Scandinavian Machine Translation)*. Uppsala, Department of Linguistics and Philology.

Tiedemann, Jörg (2001a). The plug word aligner - PWA. [Electric]. Accessed: <http://stp.ling.uu.se/corpora/plug/pwa/> [2005-05-02].

Tiedemann, Jörg (2001b). Predicting translations in context. In *Proceedings of the conference on recent advantages in natural language processing (RANLP)*, p. 240-44. Bulgaria: Tzigov Chark.

Tiedemann, Jörg (2003). *Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Diss., Uppsala University. Uppsala: Acta Universitatis Upsaliensis.

Åberg, Stina (2003). *Datoriserad analys av sammansättningar i teknisk text*. Master's thesis, Department of Linguistics and Philology, Uppsala University.

# Appendix

Here are two extracts from evaluations of link results based on lemmatized parallel corpora. An example of a partially correct link is *ensvfbell95*.

ensvfbell11:	Security measure	-	Säkerhet åtgärd	correct
ensvfbell126:	unrelated	-	inte tillhöra han släkt	2(9)
	he deal probably unrelated to	-	han släkt inte tillhöra	
ensvfbell135:	To	-	För	correct
ensvfbell137:	say	-	säga	correct
ensvfbell139:	British	-	brittisk	correct
ensvfbell139:	be	-	vara	correct
ensvfbell143:	he	-	han	2(3)
	he	-	han också	
ensvfbell154:	be speak	-	tala	correct
ensvfbell162:	have	-	ha	correct
ensvfbell178:	you	-	er	wrong
	your you	-	er ni	
ensvfbell178:	a week	-	i vecka	3(5)
	of a week	-	i vecka	
ensvfbell179:	say	-	säga	correct
ensvfbell195:	the	-	den	2(3)
	the	-	åt den	
ensvfbell195:	turn to	-	ägna åt	2(7)
	to turn the	-	med ägna åt den	
ensvfbell131:	By and by	-	Så småningom	5(5)
	and by By	-	småningom Så	
ensvfbell182:	be	-	vara	correct
ensvfbell196:	the mass	-	massa	correct
ensvfbell202:	Secretary of State	-	Utrikesminister	wrong
	State Henry Secretary of struggle	-	Henry om Utrikesminister	
ensvfbell205:	I	-	Jag	correct
ensvfbell207:	Edouard Daladier	-	Edouard Daladier	2(4)
	Daladier edouard	-	Daladier Edouard	
ensvfbell212:	the Palestine Mandate	-	Palestinamandat	3(5)
	the Palestine Mandate	-	den Palestinamandat	
ensvfbell214:	this	-	denna	correct
ensvfbell227:	Kissinger	-	Kissinger	correct
ensvfbell229:	enfranchisement	-	medborgerlig rättighet	correct
ensvfbell232:	home	-	hem	correct
ensvfbell252:	be	-	vara	correct
ensvfbell261:	where	-	där	correct
ensvfbell262:	the	-	denna	correct
ensvfbell279:	rely	-	satsa	correct
ensvfbell280:	the Middle East	-	Mellanöstern	3(5)
	Middle East the	-	Mellanöstern vars	
ensvfbell286:	population	-	folk	correct
ensvfbell286:	in turn	-	i sin tur	4(8)
	themselves in turn	-	sin vilken i tur understödja	

**Figure 1:** Extract from evaluation of the lemmatized and linked Bellow bitext

sventscan112:	innan - before	correct
sventscan235:	kunna uppstå - may result	2(5)
	kunna den uppstå  -  may the	
sventscan370:	använda - be use	correct
sventscan410:	finne - spot	2(7)
sventscan2030:	trafik+säkerhet - road safety	correct
sventscan2205:	etc - etc	correct
sventscan2234:	Utvärdering - Evaluation	correct
sventscan2305:	acceleration+kurva - the acceleration curve	3(4)
	acceleration+kurva  -  acceleration curve	
sventscan2359:	avläsning+skala - scale	correct
sventscan2361:	backe - hill	correct
sventscan3230:	smuts - line	wrong
	smuts kalk  -  dirt lime	
sventscan3924:	lagra - mount	correct
sventscan3939:	cylinder+foder - cylinder liner	correct
sventscan3944:	uppgift - information	correct
sventscan3996:	Vipparm - Rocker arm	correct
sventscan4081:	tryck+rör - the delivery pipe	3(7)
	Ta bort tryck+rör  -  Remove the delivery pipe	
sventscan5680:	undre - low	correct
sventscan5752:	packning - Gasket	correct
sventscan6170:	vätska - flow	wrong
	strömma vätska  -  flow fluid	

**Figure 2:** Extract from evaluation of the lemmatized and linked Scania bitext