# Extracting Invertible Translations from pre-aligned Texts

**Michael Carl**
Institut für Angewandte Informationsforschung,
Martin-Luther-Straße 14
66111 Saarbrücken, Germany,
carl@iai.uni-sb.de

## Abstract

This paper presents an approach to extract invertible translations from pre-aligned bilingual texts. The extracted set of invertible translations is unambiuous because each string occurs only once in either language side. Two variants of the algorithms are presented using different knowledge resources. The knowledge rich variant of the algorithm makes use of a bilingual lexicon in addition to a morphological analyser and a shallow syntax formalism which are similarly used in the knowledge poor algorithm. It is shown that the knowledge rich method yields better results than the knowledge-poor method.

## 1 Introduction

Extracting terms and term translations from bilingual texts is a mayor challange in current NPL technologies. With respect to the length of the aligned pieces of text one can distinguish between sentence alignment, phrase alignment and word alignment, even though the frontiers between these categories is sometimes difficult to define.

- Sentence alignment extracts whole sentence translations from the bilingual text. Already by means of knowledge poor methods, alignment results of more then 90% precision can be reached cf. (Macklovitch and Hannan, 1996). A recent comparision of six alignment tools has been carried out in the ARCADE project (Langlais et al., 1999). Also, sentence alignment tools are available in every commercial translation memory.
- Word alignment seeks to extract word translations from bilingual texts. Word alignment has been investigated on parallel texts (Wu and Xia, 1995; Hiemstra, 1998; Brown, 1997; Melamed, 1997) and on non-parallel texts (Fung and W, 1994). While in the former approaches the texts are assumed to be translations of each other this is not required in the latter approaches.
- Phrase alignment extracts pieces of translations between the sentence level and the word level from bilingual texts. Many approaches focus on the extraction of noun phrase translations (Jones and Alexa, 1994; Kupiec, 1993).

| Table 1 | Invertibility condition |
| --- | --- |
| $\left.\begin{array}{ccc} a & \leftrightarrow & x \\ b & \leftrightarrow & y \end{array}\right\}$ | $\Longrightarrow \left\{\begin{array}{ccc} ((a \neq b) & and & (x \neq y)) \\ & or & \\ ((a = b) & and & (x = y)) \end{array}\right.$ |

Alignment techniques have most extensively been studied in the context of cross-linguistic information retrieval (CLIR) and first of all for bilingual sentence alignment. However, automatic extraction of phrases and term translations becomes increasingly crucial as terminology grows every day and with this with the need for NLP applications and machine translation. The requirements of the type of extracted constituent may vary from application to application. While for a translation memory sentence translations are required, for text indexing and machine translation one would prefer single content words.

In this paper a method is presented which extracts from a pre-aligned bilingual text word and phrase translations. Due to its modular and rule-based components, the algorithm is suited to be quickly adapted to extract single words, such as nouns or adjectives, word clusters like NPs or entire sentences. In addition, the algorithm extracts an invertible set of translations from the pre-aligned text. The invertibility condition is particular useful to distinguish between more or less frequent collocations present in the bilingual text.

This paper first discusses the invertibility criterion. Two variants of an algorithm are presented which generate from a bilingual text an invertible set of translations. For each of the variants a small experiment is reported and the alignment results are discussed.

## 2 Invertible Bilingual Grammars

The concept of invertible grammars is not new. It has been shown that invertible grammars can be updated in polynomial time in the size of the input (Mäkinen, 1992) and that for each context-free grammar, there exists an invertible context-free grammar such that both grammars generate the same language (Harrison, 1978). As yet, to my knowledge, no research has been undertaken which applies the invertibility condition to bilingual grammars inference.

According to (Mäkinen, 1992) a (monolingual) grammar is invertible if $A \rightarrow a$ and $B \rightarrow a$ in the set of production rules implies that $A = B$. Thus, every right-hand side of a production rule occurs only once in the grammar.

Applying this condition to bilingual grammars can be paraphrased as: if any one language side of a rule is identical to another rule, then the other language side must be identical as well. A formal definition of this is given in table 1.

As an example, consider table 2 where the entries are not conform to the invertibility condition. Two pairs of translations 2.1 and 2.2 in table 2, are of the form $a \leftrightarrow x$ and $b \leftrightarrow x$. For each of the conflicting pairs of translations we would have to eliminate one - or add a disambiguating context in the right-hand sides - in order to fulfill the invertibility condition.

## 3 Invertible Alignment

An invertible translation base $TB$ bears holistic constraints on the properties of the reference text. Each translation in $TB$ has a number of features which makes it different to all other translations. There are many ways to create an invertible set from the bilingual text. For instance, provided that no two reference translations have equal left-hand or right-hand sides and different right-hand or left-hand, the reference text could serve as an invertible $TB$.

However, at least two criteria should be considered when generating an invertible $TB$: i) the extracted translations should be of type and length one is looking for and ii) it should contain the least number of erroneous or misleading translations. The resulting $TB$ should thus contain a maximum number of useful coherent translations.

Two methods are described to generate an invertible set of translation examples from a reference text. While the first method makes use of the morphological analyzer MPRO (Maas, 1996) and the shallow syntactic processor KURD (Carl et al., 1997), the second method considers in addition a bilingual lexicon. It is shown that the quality of translations is best when using the latter method.

```
Filter_Invertible_Translation_Base(AB)

usefulness_values U
exclude_list E

    For all pairs of alignments A_i, A_j ∈ AB  i ≠ j do

// check identity of A_i and A_j;
// increase usefulness for high frequency alignments

        if ((lhs_i = lhs_j) AND (rhs_i = rhs_j)) do
            U_i += W(A_j);     // increase usefulness U_i
            discard A_j from AB;
        end

// check invertibility of alignments A_i and A_j
// store index j of non-invertible A_j in E_j

        else if ((lhs_i = lhs_j) AND (rhs_i ≠ rhs_j)
               OR (lhs_i ≠ lhs_j) AND (rhs_i = rhs_j)) do
            U_i -= W(A_j);
            append j to exclude list E_i;
        end

// check compositionality of A_i and A_j;
// x ⊂ y means x is a substring of y

        else if ((lhs_i ⊂ lhs_j) AND (rhs_i ⊂ rhs_j)) do
            U_i += W(A_j);
        end

// check non-compositionality of A_i and A_j;
// decrease usefulness U_i of A_i

        else if ((lhs_i ⊂ lhs_j) AND (rhs_i ⊄ rhs_j)
               OR (lhs_i ⊄ lhs_j) AND (rhs_i ⊂ rhs_j)) do
            U_i -= W(A_j);
        end
    end    // For each pair of alignments

// generate invertible TB

    For each U_i calculate mean usefulness mean(U_i);
    Sort AB by mean(U);
    For all A_i in sorted list AB do
        include A_i in TB;
        for all j ∈ E_i do
            delete A_j from AB;
        end;
    end    // generate invertible set

end;
```

Figure 1: Filter invertible Translation Base $TB$

| Table 2 | Invertibility Clash | | |
|---|---|---|---|
| 2.1 | $\left.\begin{array}{r}\text{(Gear shift lever)}_{noun} \\ \text{(Transmission Unit Gear Selector)}_{noun}\end{array}\right\}$ | $\leftrightarrow$ | (Gestängehebel)$_{noun}$ |
| 2.2 | $\left.\begin{array}{r}\text{(Locate the outer cable)}_{s} \\ \text{(Secure the outer cable)}_{s}\end{array}\right\}$ | $\leftrightarrow$ | (Aussenseil befestigen.)$_{s}$ |

## 4 Alignment without Anchors

Both language sides of the pre-aligned reference text are morphologically analyzed with MPRO, cf. (Maas, 1996). By means of the shallow post-morphological formalis, KURD (Carl et al., 1997), a number of constituents to be extracted are marked in each language side of the reference translation. The marked constituents are nouns, adjectives, noun-clusters, verbal clusters, simple NPs, DPs and PPs. Since there is no knowledge which left-hand side constituent pairs with which right-hand side constituent all pairing combinations are extracted as potential alignments. Thus, if $m$ constituents are marked in the left-hand side of a reference translation and $n$ constituents are marked in the right-hand side, $m \times n$ alignments are extracted and added to an alignment base $AB$. Each alignment $A \in AB$ is weighted according to the difference of the number of words in their right-hand sides (rhs) and left-hand sides (lhs) as shown in equation 1.

$$W(A : lhs \leftrightarrow rhs) \quad = \quad \frac{min(lhs, rhs)}{max(lhs, rhs)} \qquad (1)$$

Alignments where the number of words in lhs is equal to the number of words in the rhs have the weight 1. The more the number of words differ in both language sides the smaler is the wight of the alignment.

Once the entire reference text is treated this way, an invertible set is filtered from the extracted alignments. A usefulness value $U$ is computed for each alignment based on the number of excluding and reinforcing alignments. The algorithm is shown in Figure 1. For each pair of alignments it is checked whether they conform to the invertibility condition. In case two alignments violate the invertibility condition their usefulness value is mutually decreased by the weight $W$ of the conflicting alignment. Otherwise the compositionality of the alignments is checked. In case one alignment is contained in the other one, their usefulness value are mutually augmented. In case one hand side of an alignment is contained in the other alignment but the other side isn't this alignment pair is non-compositional and their usefulness is mutually decreased.

Once all pairs of alignments are checked, the mean usefulness for each alignment is calculated. The most 'useful' alignment is, then, included into the initial $TB$ and all conflicting alignments are discarded from the $AB$. Then the next remaining, most useful alignment is added to $TB$ and its conflicting alignments are excluded from $AB$ and so forth until no more alignments are in $AB$. The algorithm is depicted in Figure 1

## 5 An Alignment Example

A reference text consisting of 13 German-English reference translations as depicted in table 3 was given to the system. The aim was to find possible translations of German *stark*. In the reference translations *stark* translates into English *strong, big, high, heavy, bad, grave, best, large* and *considerable.* In four of the reference translations *stark* translates into *strong* and once into *strongly*. The translation *stark* $\leftrightarrow$ *strong* would thus be a potential candidate for a default translation to be found by the system. The remaining eight translations of *stark* occur only once. They would need to occur with some context in order to be unique and thus conform to the invertibility condition.

349 alignments were extracted including the original 13 reference translations. There were 6 redundant alignments i.e. alignments that occur twice or more. These were due to the fact that for alignments to be identical the lemma (stem) and its part of speech are considered. However, because words may have ambiguous interpretations, the same sequence can be once extracted as a a noun and once as an adjective or it can be tagged as a `np` and by another rule as a `noun`. From the 349 alignments, 55 translations were retained.

A subset of the retained invertible translations containing the word *stark* in their lhs is shown in table 4. For the sake of better reading the word's surface forms have been used instead of their lemma. 12 of the 13 collocations of *stark* have been correctly found. There are two entries in the table having as their lhs *stark* which apparently contradict the invertibility condition. However, this is only on the surface so since the translation *stark* $\leftrightarrow$ *strongly* is the adverb translation while *stark* $\leftrightarrow$ *smoker* is the

| Table 3 | | Reference translations containing *stark* |
|---|---|---|
| Das ist ein **starker** Mann | $\leftrightarrow$ | This is a **strong** man |
| Es war sein **stärkstes** Theaterstück | $\leftrightarrow$ | It has been his **best** play |
| Wir Hoffen auf eine **starke** Beteiligung | $\leftrightarrow$ | We hope a **large** number of people will take part |
| Eine 100 Mann **starke** Truppe | $\leftrightarrow$ | A 100 **strong** unit |
| Der **starke** Regen überraschte uns | $\leftrightarrow$ | We were surprised by the **strong** rain |
| Maria hat **starkes** Interesse gezeigt | $\leftrightarrow$ | Mary has shown **strong** interest |
| Paul hat **starkes** Fieber | $\leftrightarrow$ | Paul has **high** temperature |
| Das Auto war **stark** beschädigt | $\leftrightarrow$ | The car was **badly** damaged |
| Das Stück fand einen **starken** Widerhall | $\leftrightarrow$ | The piece had a **considerable** response |
| Das Essen war **stark** gewürzt | $\leftrightarrow$ | The meal was **strongly** seasoned |
| Hans ist ein **starker** Raucher | $\leftrightarrow$ | John is a **heavy** smoker |
| Im Sommer gab es eine **starke** Nachfrage | $\leftrightarrow$ | There was a **big** demand in summer |
| Er hatte daran **starken** Zweifel | $\leftrightarrow$ | He had **grave** doubts on it |

| Table 4 | | Invertible translations containing *stark* |
|---|---|---|
| $(stark)_{\mathtt{adv}}$ | $\leftrightarrow$ | $(strongly)_{\mathtt{adv}}$ |
| $(stark)_{\mathtt{adj}}$ | $\leftrightarrow$ | $(smoker)_{\mathtt{noun}}$ |
| $(starker\ Zweifel)_{\mathtt{np}}$ | $\leftrightarrow$ | $(grave\ doubt)_{\mathtt{np}}$ |
| $(stark\ beschädigt)_{\mathtt{verb}}$ | $\leftrightarrow$ | $(badly\ damaged)_{\mathtt{verb}}$ |
| $(stark\ gewürzt)_{\mathtt{np}}$ | $\leftrightarrow$ | $(strongly\ seasoned)_{\mathtt{np}}$ |
| $(starke\ Truppe)_{\mathtt{np}}$ | $\leftrightarrow$ | $(strong\ unit)_{\mathtt{np}}$ |
| $(starker\ Mann)_{\mathtt{np}}$ | $\leftrightarrow$ | $(a\ strong\ man)_{\mathtt{dp}}$ |
| $(ein\ starker\ Mann)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(strong\ man)_{\mathtt{np}}$ |
| $(stärkstes\ Theaterstück)_{\mathtt{np}}$ | $\leftrightarrow$ | $(best\ play)_{\mathtt{np}}$ |
| $(starkes\ Interesse)_{\mathtt{np}}$ | $\leftrightarrow$ | $(strong\ interest)_{\mathtt{np}}$ |
| $(starkes\ Fieber)_{\mathtt{np}}$ | $\leftrightarrow$ | $(high\ temperature)_{\mathtt{np}}$ |
| $(starker\ Regen)_{\mathtt{np}}$ | $\leftrightarrow$ | $(strong\ rain)_{\mathtt{np}}$ |
| $(der\ starke\ Regen)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(the\ strong\ rain)_{\mathtt{dp}}$ |
| $(starker\ Raucher)_{\mathtt{np}}$ | $\leftrightarrow$ | $(a\ heavy\ smoker)_{\mathtt{dp}}$ |
| $(ein\ starker\ Raucher)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(heavy\ smoker)_{\mathtt{np}}$ |
| $(starken\ Widerhall)_{\mathtt{np}}$ | $\leftrightarrow$ | $(considerable\ response)_{\mathtt{np}}$ |
| $(einen\ starken\ Widerhall)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(a\ considerable\ response)_{\mathtt{dp}}$ |
| $(starke\ Nachfrage)_{\mathtt{np}}$ | $\leftrightarrow$ | $(big\ demand)_{\mathtt{np}}$ |
| $(eine\ starke\ Nachfrage)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(a\ big\ demand)_{\mathtt{dp}}$ |
| $(starke\ Beteiligung)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(will\ take)_{\mathtt{verb}}$ |
| $(eine\ starke\ Beteiligung)_{\mathtt{dp}}$ | $\leftrightarrow$ | $(people\ will\ take\ part)_{\mathtt{np}}$ |
| $(auf\ ein\ starke\ Beteiligung)_{\mathtt{pp}}$ | $\leftrightarrow$ | $(of\ people\ will\ take\ part)_{\mathtt{pp}}$ |

adjective translation. Instead of *strong*, the noun *smoker* was chosen as the default adjective translation.

There are as well other noisy translations such as *Maria* $\leftrightarrow$ *interest*, *Raucher* $\leftrightarrow$ *John*, *Regen* $\leftrightarrow$*by the strong rain* among the 55 extracted translations. One reference translation is decomposed in a completely misleading way as the translation *Beteiligung* $\longleftrightarrow$ *take*, *starke Beteiligung* $\longleftrightarrow$ *will take* etc. implies.

However, the result is encouraging as 2/3 of the extracted translations are correct. Moreover, erroneous translations occur mostly for single frequency words such that one can expect reasonable translations to be extracted as the size of the reference text grows and word occurrences increase. The most reasonable adverb-default translation *stark* $\leftrightarrow$ *strongly* has been found and for almost all of the less frequent translations of *stark* the context is included in the translation. To give the system more security over its decisions the bilingual lexicon has been included in a second alignment experiment.

## 6 Alignment with Lexical Anchors

In a second experiment lexical anchors were used to calculate the alignment weight. Similar to the previ-

| Table 5 | Bilingual Lexicon | |
|---|---|---|
| $(stark)_{\mathtt{adv}}$ | $\longleftrightarrow$ | $(strong)_{\mathtt{adv}}$ |
| $(stark)_{\mathtt{adv}}$ | $\longleftrightarrow$ | $(bad)_{\mathtt{adv}}$ |
| $(stark)_{\mathtt{adv}}$ | $\longleftrightarrow$ | $(good)_{\mathtt{adv}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(strong)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(big)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(high)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(heavy)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(grave)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(large)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(considerable)_{\mathtt{adj}}$ |

| Table 6 | Anchored Invertible Translation Examples | |
|---|---|---|
| $(stark)_{\mathtt{adj}}$ | $\longleftrightarrow$ | $(strong)_{\mathtt{adj}}$ |
| $(stark)_{\mathtt{adv}}$ | $\longleftrightarrow$ | $(strong)_{\mathtt{adv}}$ |
| $(starker\ Mann)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(strong\ man)_{\mathrm{np}}$ |
| $(ein\ starker\ Mann)_{\mathrm{dp}}$ | $\longleftrightarrow$ | $(a\ strong\ man)_{\mathrm{dp}}$ |
| $(starke\ Truppe)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(strong\ unit)_{\mathrm{np}}$ |
| $(starke\ Zweifel)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(grave\ doubts)_{\mathrm{np}}$ |
| $(starke\ Regen)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(strong\ rain)_{\mathrm{np}}$ |
| $(starkes\ Interesse)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(strong\ interest)_{\mathrm{np}}$ |
| $(starkes\ Fieber)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(high\ temperature)_{\mathrm{np}}$ |
| $(stark\ beschädigt)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(badly\ damaged)_{\mathrm{np}}$ |
| $(starken\ Widerhall)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(considerable\ response)_{\mathrm{np}}$ |
| $(einen\ starken\ Widerhall)_{\mathrm{dp}}$ | $\longleftrightarrow$ | $(a\ considerable\ response)_{\mathrm{dp}}$ |
| $(stark\ gewürzt)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(strongly\ seasoned)_{\mathrm{np}}$ |
| $(starker\ Raucher)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(heavy\ smoker)_{\mathrm{np}}$ |
| $(ein\ starker\ Raucher)_{\mathrm{dp}}$ | $\longleftrightarrow$ | $(a\ heavy\ smoker)_{\mathrm{dp}}$ |
| $(starke\ Nachfrage)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(big\ demand)_{\mathrm{np}}$ |
| $(eine\ starke\ Nachfrage)_{\mathrm{dp}}$ | $\longleftrightarrow$ | $(a\ big\ demand)_{\mathrm{dp}}$ |
| $(starke\ Beteiligung)_{\mathrm{np}}$ | $\longleftrightarrow$ | $(large)_{\mathtt{adj}}$ |
| $(eine\ starke\ Beteiligung)_{\mathrm{dp}}$ | $\longleftrightarrow$ | $(large\ number)_{\mathrm{np}}$ |
| $(auf\ eine\ starke\ Beteiligung)_{\mathrm{pp}}$ | $\longleftrightarrow$ | $(a\ large\ number)_{\mathrm{dp}}$ |

ous algorithm, in a first step possible alignment candidates are extracted from the reference text and in a second step an invertible set $TB$ is filtered from the alignments. In contrast to the previous approach, only alignments which contain at least one lexical anchor are extracted.

The alignment weight $W$ - as shown in equation 2 - is computed based on the difference in length of the lh and rh-sides of the alignments and the sum of translation probabilities $P(lhs_{k,l}, rhs_{o,p})$ of the lexical anchors.

$$W(A : lhs \leftrightarrow rhs) = \frac{min(lhs, rhs) * L(lhs, rhs)}{max(lhs, rhs)} \quad (2)$$

$$L(lhs_{i,m}, rhs_{j,n}) = \sum_{k<l} \sum_{o<p} P(lhs_{k,l}, rhs_{o,p}) \quad (3)$$

The bilingual lexicon shown in table 5 contains 10 entries; each of which has the word *stark* in its lhs. Each of these translations has a probability of 0.1. The last alignment in table 6 has 4 words in its lhs and 3 words in its rhs. According to equation 2, the wight $W$ of this alignment is thus $3 * 0.1/4 = 0.075$. For a small experiment the reference text shown in table 3 has been used. In adition, the bilingual lexicon in table 5 has been used containing a couple of lexical translations rules for *stark*.

108 alignments were extracted from the reference text from which 37 translations were kept in the $TB$.

In table 6 the extracted translations which contain *stark* in their lh-sides are depicted. All collocations have been found within a minimal disambiguating context. The translation *starke beteiligung* $\leftrightarrow$ *large* and consequently *eine starke Beteiligung* $\leftrightarrow$ *large* *number* seems somewhat odd[1]. For both, the adjective and the adverb interpretation the reasonable translation *stark* $\leftrightarrow$ *stong* and *stark* $\leftrightarrow$ *strongly* are generated. Almost 90% of the translations are correct; a single lexical anchor, thus, yields sufficient indices to determine the surrounding constituents.

## 7 Conclusion

This paper presents an algorithm which extracts an invertible set of translations from a pre-aligned bilingual text. The invertibility condition requires that each string in either side of a translation occurs at most once in the whole set. Instead of allowing translation ambiguities, each translation contains a desambiguating context which makes it unique in the set.

The algorithm works in two steps. First, potential alignment candidates are extracted from the bilingual text. In a second step, the most useful invertible set is filtered from the alignment candidates. The usefulness of a translation is calculated based on its frequency, its compatibility in the aligned set and its compositionality.

Two versions of the algotithm are presented; one version makes use of a morphological analyser and a shallow syntactic formalism. The syntactic formalism serves to detect and mark those types of constituents that are to be extracted from the bilingual text. In addition to this, a second version uses a bilingual lexicon to anchor the marked constituents.

Both versions of the alignment algorithm are evaluated on a small artificial bilingual text. Unsurprisingly the knowledge richer method using a bilingual lexicon yields better results. The algorithm can be quickly adapted to extract different types of constituent by simply replacing or modifying the set of rules which mark the translation sought for.

## References


Ralf D. Brown. 1997. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *TMI-97*, pages 111–118.

Michael Carl, Antje Schmidt-Wigger, and Munpyo Hong. 1997. KURD - a Formalism for Shallow Postmorphological Processing. In *Proceedings of the NLPRS*, Phuket, Thailand.

Pascale Fung and u Dekai W. 1994. K-vec: A new approach for aligning parallel texts. In *COLING-94*.

Michael A. Harrison. 1978. *Introduction to Formal Language Theory*. Addison-Wesley, Reading.

Djoerd Hiemstra. 1998. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *CLIN*, pages 41–58.

Daniel Jones and Melina Alexa. 1994. Towards Automatically Aligning German Compound with English Word Groups in Example-Based Translation. In *NeMLaP*.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the ACL*.

Philippe Langlais, Michel Simard, and Jean Véronis. 1999. ARCADE Methods and Practical Issues in Evaluating Alignment Techniques. Technical report, University of Aix, http://www.lpl.univ-aix.fr/projects/arcade/report1-en/.

Heinz-Dieter Maas. 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser, editor, *Linguistische Verifikation, Sprache und Information*. Max Niemeyer Verlag, Tübingen.

Elliott Macklovitch and Marie Louise Hannan. 1996. Line 'EM UP: Advances in alignment technology and their impact on translation support tools. In *Proceedings of the AMTA*, Montreal.

Erkki Mäkinen. 1992. On the structural grammatical inference problem for some classes of context-free grammars. *Information Processing Letters*, 42:1–5.

I.D. Melamed. 1997. A word-to-word model of translation equivalence. In *35th Conference ACL*.

D. Wu and X. Xia. 1995. Large-scale automatic extraction of an english-chinese translation lexicon. *Machine Translation*, 9:285–313.


[1] There was no rule which marked the phrase *large number of people* such that an alignment containing this translation could not be generated.