A truly multilingual, high coverage, accurate, yet simple, subsentential alignment method

Adrien Lardilleux and **Yves Lepage**
GREYC, Université de Caen Basse-Normandie,
BP 5186, Caen Cedex, France
Firstname.Lastname@info.unicaen.fr

Abstract

This paper describes a new alignment method that extracts high quality multi-word alignments from sentence-aligned multilingual parallel corpora. The method can handle several languages at once. The phrase tables obtained by the method have a comparable accuracy and a higher coverage than those obtained by current methods. They are also obtained much faster.

1 Introduction

Alignment is an important task in natural language processing for a variety of purposes like the constitution of lexical resources, machine translation or cross-lingual information retrieval. In the case of machine translation, alignment serves as a starting point to generate phrase tables (Koehn et al., 2003), which are the primary source of knowledge for most data-driven machine translation systems.

Several alignment tools are freely available today. Among them, the bilingual phrase aligner Giza++ (Och and Ney, 2003) can perform high quality alignments based on words statistics. It is considered the most efficient tool. Some criticisms may however be addressed to this kind of tool.

Firstly, numerous parameters have to be tuned in order to optimize the results for a particular alignment task, which can be very time consuming. This is all the more important when multilingual alignment is concerned, since every language pair will require its own set of parameter values. For this reason, Moore (2005) proposes an alternative to Giza++, and reports an accuracy close to the probabilistic generative approach (Giza++), yet using a simpler — and faster — method.

Secondly, freely available tools cannot align a large number of languages simultaneously. They can only handle pairs of languages. This results in a complexity explosion when multilingual alignments are needed. Giguet and Luquet (2006) report alignments in 20 languages, but only pairs between English and 19 other European languages are considered. Simard (1999) showed how to adapt a bilingual method to align more than two versions of a text at the sentence level, but this requires to identify first which of the language pairs are the most "similar": languages still require to be processed by pairs.

We propose an approach that differs from all previous techniques. It is intended to be multilingual, fast, simple, yet accurate. Following Marcu and Wong (2002), phrase alignments can directly be obtained without any intermediate word alignment step. And more generally, any sequence of words can be obtained. The method relies on simple heuristics based on similarities and differences between sentences, such as used by Cicekli (2000).

The paper is organized as follows. Section 2 gives an overview of the basic concepts used in the proposed multilingual alignment technique. Section 3 describes the technique in more details. Section 4 shows some alignment results on actual data. Section 5 elaborates on some possible optimization. Section 6 compares the method with state-of-the-art tools.

2 Simple is beautiful

2.1 "Perfect" alignments

Perfect alignments may be evidenced using any similarity coefficient like the cosine similarity, Jaccard index or Dice coefficient. They are alignments that get a maximal score of 1 according to any of these similarity coefficients. They contain those words that appear exactly on the same lines.

Consider for instance the following bilingual toy corpus, where each line is a pair of aligned sentences (source words are in lowercase and target words in uppercase):

$$\begin{array}{c} a\ d \leftrightarrow A\ D \\ b \leftrightarrow B \\ b \leftrightarrow C \\ a\ e \leftrightarrow A\ D\ D \end{array}$$

In this toy corpus, a and A appear exactly on the same lines, with the same number of occurences. Consequently, there are good chances that they be lexical equivalences. In addition, we can extract the contexts of a and A: from the first line, we can say that d is likely to be an equivalent of D, and from the last line, that e is likely to be an equivalent of $D\ D$.

2.2 Forging perfect alignments by corpus splitting

The previous considerations fail in aligning the source word b, because it translates to B on the second line and to C on the third one. An alignment method should deliver scores that reflect the probability of a given target word to be a translation of a source word.

To answer this requirement, we split the corpus into two subcorpora in the following manner. Then we look for perfect alignments in each of the subcorpora:

$$\begin{array}{c} a\ d \leftrightarrow A\ D \\ b \leftrightarrow B \\ \hline b \leftrightarrow C \\ a\ e \leftrightarrow A\ D\ D \end{array}$$

b can now be "perfectly" aligned with B in the first subcorpus, and with C in the second one. As a result, when considering all possible alignments for b, we can say that b translates to B with an observed probability of 0.5 and to C with an observed probability of 0.5. In the other direction, B translates to b with an observed probability of 1 and C translates to b with an observed probability of 1.

2.3 Segmentation as an alignment process

None of the operations mentioned so far is restricted to language pairs. In fact, the previous examples could run with absolutely no change whatever the number of languages in which the sentence-aligned corpus is available.

We can go even further by completely striking down boundaries between languages: assimilating a multilingual corpus to a monolingual corpus. All we need to find is a set of words that strictly appear on the same lines. The language they belong to has no importance: they can all come from a different language, or all from the same one. In the second case, we've just found collocations (possibly with a low frequency).

As a result, the method we propose naturally unifies segmentation and alignment steps. The output of the method ranges from single words to complete sentences and more generally any sequence of words from existing sentences. They are monolingual collocations or bilingual, trilingual, ..., multilingual alignments.

3 A sketch of the method

In the following, we consider a corpus C in L languages. If $L = 1$, then the corpus is monolingual. The corpus is made of N lines. A line is made of L aligned utterances, one per language.

We define a multilingual alignment as a multilingual ordered sequence of words. Each line in the corpus is thus a multilingual alignment on its own right. Trivially, because of the ordering, any word in language l ($1 \leq l \leq L$) always appears before any word of language $l + 1$ in any multilingual alignment.

3.1 Splitting the corpus into subcorpora of any size

In Section 2.2, we saw that the method relies on corpus splitting to extract only "perfect" alignments. Theoretically, the method should be applied on all possible subcorpora of C. This is not feasible in practice, because there are 2^N possible subcorpora.[1]

The answer to the large number of subcorpora is sampling. This is done iteratively, with various sample sizes. More precisely, the full coverage of the corpus is ensured by partitioning. This permits fast processing while maintaining the subcorpora representative of the initial corpus as much as possible.

3.2 Extracting alignments

At each iteration, the initial corpus C is randomly partitioned into M subcorpora of n lines (see Section 5 for a discussion about the possible values for n) and one with the remainder of lines ($M \times n + r = N$), leaving the content of the lines unchanged. Iterations are independent.

For each subcorpus obtained by partitioning, we proceed as follows:

1. make groups of words according to the lines they appear on: each group is made of words that appear exactly on the same lines of the subcorpus (“perfect” alignments);
2. for each group, go through the lines it appears on. Two sequences of words are extracted from each line:
   (a) the group of words;
   (b) all the context of the group of words on the line.

   We say that (a) delivers direct sequences, and (b) context sequences. As word order is preserved in both extracted sequences, they constitute multilingual alignments.

Any alignment may be obtained a plurality of times, from different iterations, different subcorpora and different lines. The result of the process is a list of alignments with the count of the number of times they have been obtained.

In the general case, the method outputs noncontiguous sequences of words. They can subsequently be filtered according to specific criteria, like word contiguity, number of languages covered, or the number of words in a given language.

3.3 Computing one-to-many translation probabilities

As sketched above (Section 2.2), the observed probability that a sequence of words S_i in language i translates into S_j in language j is the number of alignments that exactly contain both sequences in the respective languages, $C(S_j, S_i)$, over the total number of alignments where S_i appears, $C(S_i)$:

$$P(S_j|S_i) = \frac{C(S_j, S_i)}{C(S_i)}$$

Each alignment is weighted according to the number of times it is obtained during the alignment generation process.

This is similar to the way Koehn et al. (2003) estimate phrase translation probabilities. Consequently, the proposed technique outputs translation tables directly usable by statistical machine translation software.

3.4 Implementation

A single Python script implements the previously described method. It is open source and available at `http://users.info.unicaen.fr/~alardill/malign/`.

In the following, iterations are processed sequentially. Execution times are measured on a machine equipped with a 2.2 GHz processor, using the Psyco[2] JIT compiler to speed up code execution.

4 Testing the method

4.1 Data used in the subsequent experiments

We used the English, Japanese and Arabic training parts of the IWSLT 2007 machine translation campaign corpus (Fordyce, 2007) to conduct our experiments. This is nearly 20,000 triples of aligned sentences from the BTEC (Basic Traveler Expression Corpus) (Takezawa et al., 2002). The corpus has been tokenized and lowercased for the English part. Figure 1 shows an excerpt of the data.

4.2 First experiment: aligning three languages at a time

In a first experiment, we set the number of iterations to 100 and use new random sizes of subcorpora at

[1] The number of subcorpora of size p for a corpus of size N is $\binom{N}{p}$. The total number of subcorpora is $\sum_{p=1}^{N} \binom{N}{p} = 2^N$.

[2] `http://psyco.sourceforge.net/`

هل هناك مكتبة أخرى تبيع ها ؟ */hl hnāk mktbh ʾaḫrā tbyʿ hā ?/*	↔	does another bookstore sell it ?	↔	別の書店で売ってますか。 */betu no syoten de u te masu ka ./*
أود فوطة . */ʾawd fwṭh ./*	↔	i ’d like a towel .	↔	タオルが欲しいのですが。 */taoru ga hosii no desu ga ./*
لكل حاجته . أنا آخذ جعة . */lkl ḥāǧth . ʾanā ʾāḫḏ ǧʿh ./*	↔	to each his own . i ’m having a beer .	↔	人それぞれね。私はビールにする。 */hito sorezore ne . watasi ha biiru ni suru ./*

Figure 1: Excerpt of the data used in the experiments. Each line is a triple of aligned utterances in Arabic, English and Japanese. Transliterations are not part of the original corpus.

Arabic	English	Japanese	Freq.
	beer	ビール	202
بيرة	beer	ビール	35
بيرة	beer		14
البيرة	beer	ビール	8
البيرة	beer		8
جعة	beer	ビール	3
⋮	⋮	⋮	⋮

Table 1: Most frequent multilingual alignments obtained with “beer” as the English sequence. These alignments are used to extract the translation of “beer” into Arabic and Japanese.

each iteration. We keep only contiguous sequences of words in each language and keep only those alignments containing a non-empty sequence in at least 2 languages. The whole process took 94 seconds. It yielded 116,944 unique multilingual alignments. The sum of all absolute frequencies is 917,532. The average number of times an alignment was obtained in this experiment is thus 7.8 times.

Table 1 shows the most frequent multilingual alignments with “beer” as the English sequence, along with their frequencies. The complete table contains 19 alignments. The sum of their frequencies is 290. From the complete table, we determine the probability of each translation of “beer” in Arabic and Japanese according to the formula given in Section 3.3.

The best translations for “beer” can be seen on Table 2. Both tables were simultaneously obtained in one pass. The tables show that the translations obtained are indeed correct translations, those with low probabilities being noise. Further experiments have shown that this noise does not alter the quality of subsequent machine translation tasks.

The same process can apply to all English sequences of words to produce an English-to-Arabic set of alignments along with their translation probabilities (the same for English-to-Japanese). By doing this once for each language, all pairs of languages are covered in both directions (source to target and target to source). In other words, we obtain all possible bilingual translation tables (quadratic in the number of languages) in linear time. On our data, this was done in 3 passes (3 languages), and took 20 seconds for more than 100,000 alignments.

4.3 Second experiment: extracting syntactic patterns

A feature of the method is that it extracts monolingual syntactic patterns. Such syntactic patterns can be obtained by applying the method on a monolingual corpus, such as the English part of the preceding corpus, and retaining those “alignments” that contain at least two words (including discontinuous sequences of words).

252,848 unique “monolingual alignments” were obtained on the previous data, the cumulated frequencies of which equals 18,544,963. The most frequent patterns are presented on Table 3, along with their frequencies. Patterns made up of repeated words only, such as “the the” (62,337 in total) are not shown.

An important characteristic is that the most frequent alignments are not necessarily the most fre-

Arabic			Prob.
بيرة	/*byrh*/	‘beer’	0.593
البيرة	/*ālbyrh*/	‘a beer’	0.186
جعة	/*ǧʿh*/	‘beer’	0.047
البيرة المحلية	/*ālbyrh ālmḥlyh*/	‘local beer’	0.035

Japanese			Prob.
ビール	/*biiru*/	‘beer’	0.970
国産 ビール	/*kokusan biiru*/	‘local beer’	0.019
缶	/*kan*/	‘can (tin)’	0.004
しか	/*sika*/	‘only’	0.003

Table 2: The four first Arabic and Japanese translation candidates with highest scores for “beer”, extracted from multilingual alignments (see Table 1). Notice the difference in scores between the first candidate and the other ones.

English	Freq.
i ’d like some .	6,786
i ’d like a .	6,405
, please .	6,215
where is the ?	4,247
it ’s .	4,170
los angeles	3,609
do you have ?	2,367
where ’s the ?	2,314
is this ?	2,055
do you have any ?	2,013

Table 3: The most frequent syntactic patterns obtained by applying the proposed alignment technique on the English part of the corpus only.

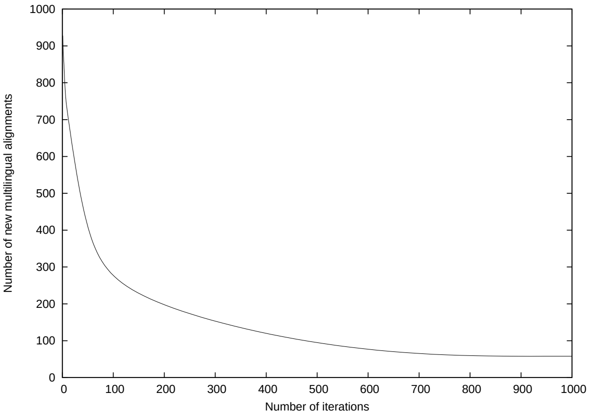

Figure 2: Number of new multilingual alignments obtained according to the number of iterations, using our trilingual corpus. The more iterations, the less new alignments. Random sizes of subcorpora were used at each iteration.

quent patterns. All syntactic patterns, whatever their frequencies, can be extracted in a similar way with this approach. Multilingual alignments of syntactic patterns are naturally obtained by the method, as it processes several languages in parallel by construction.

5 Improvement of speed and coverage

As seen previously, the iterative random partitioning process does not ensure that all possible alignments will be generated. Rather, it produces a set of alignments at each iteration, which is merged with the results of previous iterations. Some alignments may already have been produced in previous iterations, in which case it is just a matter of updating their frequency. The more iterations, the more new alignments, but the number of new alignments obtained at each iteration decreases, as shown in Figure 2.

In the following, we propose to determine which sizes of subcorpora have to be favored when partitioning the corpus, in order to improve results in a typical machine translation task. Here, we make use of the complete Japanese and English parts of the previously mentioned corpus. This is roughly 40,000 lines in total.

5.1 More alignments with smaller subcorpora

As we reach for efficiency, we look for a maximum number of alignments in as few iterations as possible. The reason for doing so is that, typically, in statistical machine translation, more alignments lead to better translation quality.

In a first experiment, we determine which sizes of subcorpora produce the highest number of alignments. The results are shown on Figure 3. Smaller subcorpora yield significantly more alignments than larger ones.[3]

[3] Note that subcorpora of size 1 and subcorpora of size N (here, $N = 40,000$) always output the same alignments, what-

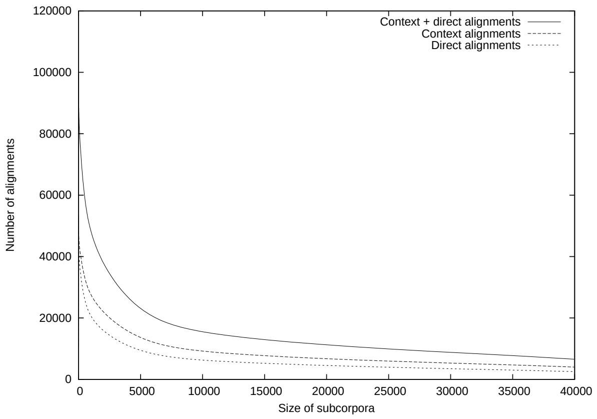

Figure 3: Number of alignments obtained when running 1 iteration according to the size of subcorpora. The smaller the subcorpora, the more alignments.

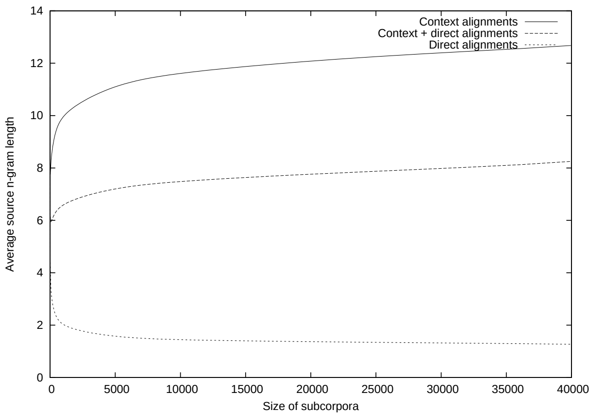

Figure 4: Average source n-gram length against the size of the subcorpora they were extracted from. The "direct" alignments are short for large subcorpora, while the contexts are small for small subcorpora, and vice versa. In total, the whole set of alignments follows the contexts' behavior because they are more numerous.

If the subcorpora size selection favored smaller subcopora, much more alignments would be obtained in no time. However, exhaustivity of these alignments is not guaranteed, because neglecting larger subcorpora may prevent some alignments to be produced. This issue is investigated in the next sections.

5.2 Shorter alignments with smaller subcorpora

In a second experiment, we measure the length of the sequences of words that appear in an alignment according to the size of the subcorpus it comes from. For subcorpora of smaller sizes (*e.g.*, down to a single line), almost all words having the same distribution, they typically end up in the only possible alignment: the original pair of sentences itself. Partitioning the 40,000 lines of the previously mentioned corpus into subcorpora of length 1 yields roughly 40,000 multilingual alignments of maximal length: all sentences are output without any change.[4]

As a result, one could expect that the larger the subcorpora, the smaller the units aligned. This is true to some extent only, as shown in Figure 4. Indeed, the method also produces alignments by context (see Section 3.2).

Context alignments are typically slightly more numerous than the "direct" ones, whatever the size of the subcorpora (see Section 5.1). The former have greater impact in all subsequent processes. Consequently, on the whole, most short sequences of word are obtained with smaller subcorpora.

5.3 Do we really need large subcorpora?

To confirm the previous insights, we determine which subcorpora produce the highest number of alignments that match an entry in a bilingual Japanese-English dictionary.[5] The reason for doing so is that a dictionary generally contains the smallest sequences of words that should be aligned, typically single words.

The results are shown on Figure 5. The maximum number of alignments found in the dictionary corresponds to smaller subcorpora (less than 1,000 lines).

5.4 Subcorpora size selection strategy

The conclusion of the three previous studies is that we should favor smaller subcorpora, *i.e.*, bias our splitting method towards smaller subcorpora, because (1) they produce more alignments (Sec-

ever the number of iterations, because there is only one way to partition the original corpus using these sizes.

[4] Except those in which a word appears several time: the sentence $a\ b\ a$ may never be output as it is, because the frequencies of a and b are different: $a\ a$ and b are highly probable to be extracted separately.

[5] We used the EDICT English-Japanese dictionary (115,000 entries): `http://www.csse.monash.edu.au/~jwb/j_edict.html`.

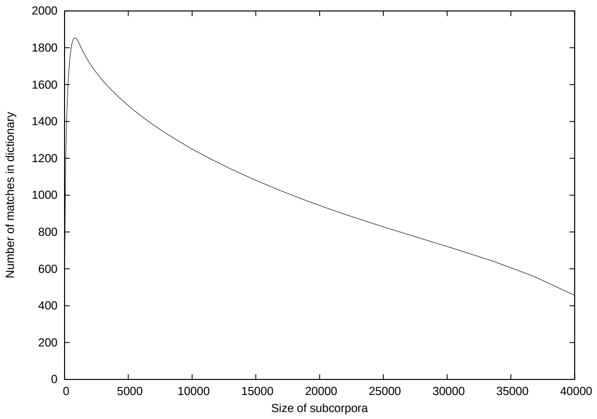

Figure 5: Number of alignments found in a bilingual dictionary according to the size of the subcorpora they were extracted from. The maximum is obtained for subcorpora made of only 500 lines in this experiment.

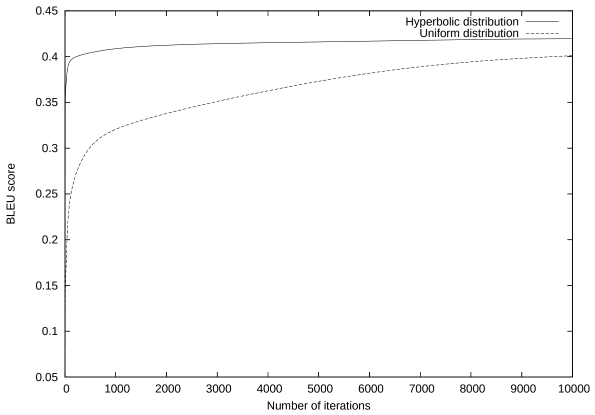

Figure 6: More iterations lead to better translation quality. Better translations are obtained faster using small sizes of subcorpora. Consequently, a hyperbolic distribution is much more efficient than a uniform one for the choice of the sizes of subcorpora.

tion 5.1), (2) they produce short sequences of words (Section 5.3), (3) while producing longer sequences as well (Section 5.2).

To confirm this, we evaluate a phrase-based SMT system on the IWSLT'07 Japanese to English classical task with two subcorpora size selection strategies: a uniform distribution (all sizes have the same probability to be chosen) and a hyperbolic distribution (the probability is inversely proportional to the size: small sizes are favored).

The test set consists in roughly 500 Japanese utterences. We use the Moses decoder (Koehn et al., 2007) to perform this experiment. Several runs are performed using phrase tables generated from alignments obtained with different numbers of iterations. Each entry in the phrase tables contains a source and a target sequences of words, as well as two translation probabilities (target to source and source to target). Note that no tuning is performed before translation; more precisely, the same parameter values from a previous tuning is reused for all runs. The evaluation uses the BLEU metric (Papineni et al., 2002).

The results are plotted on Figure 6. The more iterations, the higher the score in both cases. The score increases much faster with the hyperbolic distribution. The uniform distribution never reaches the results of the hyperbolic distribution, whatever the number of iterations (even greater than 10,000 — not plotted), although they both tend to the same score asymptotically. This confirms the conclusion that favoring shorter subcorpora yields more useful alignments.

6 Comparing the optimized method with state-of-the-art tools

The absence of tuning before decoding typically yields lower scores. In an experiment, proper tuning and a phrase table obtained with 10,000 iterations of the proposed method yields a BLEU score of 0.45 on the previous task. The same score is achieved with Moses' default phrase tables (refined alignments from IBM model 4 with lexical weighting).

As for coverage, phrase tables generated with the proposed method may have various sizes, depending on the number of iterations run. Typically, a phrase table leading to reasonable translation quality is at least twice as big as Moses' default: extra multi-word units are aligned. Those with a low score have no influence on subsequent processes.

For a more persuasive experiment, we evaluate the number of words present in the training corpus but absent from the phrase tables. Practically, these are not unseen words, but they are considered as such because the alignment method failed to extract them. As a result, the decoder is unable to translate them.

On the previous Japanese-English translation task, 3,849 words were present in the Japanese part of the training corpus but absent in Moses' default phrase tables. With the proposed method, only 322 words were not present in the phrase table. This coverage is much higher than those obtained with refined alignments (10 times less of "unknown words"). Consequently, more words are translated during decoding: 335 words are not translated using default phrase tables, while only 43 are not translated with the proposed method.

The reason why the proposed method has a higher coverage is that a given pair of aligned sentences can produce as many different alignments for a sequence of words as there are subcorpora in which this sequence occurs, while most traditional approaches yield one. For instance, assume we have the following pair of English-French sentences:

Je voudrais de la bière . ↔ I 'd like beer .

In such a situation, IBM model 4 would typically align "de la bière" with "beer", which is correct in this context. The method we propose is likely to produce the same alignment, plus other ones like "bière" with "beer", or "la bière" with "beer", depending on the subcorpora. This can be achieved without the need for extra examples where these translation links would explicitly appear.

7 Conclusion

We introduced a new multi-word alignment technique.

It has been shown to be simple: the method consists in splitting a corpus into subcorpora of any size and to look for perfect alignments and their contexts in the subcorpora.

It is truly multilingual: any number of languages can be processed at the same time.

It is accurate: in some experiment, it matched the accuracy of refined alignments obtained from IBM model 4, while exhibiting a much higher coverage.

References

Ilyas Cicekli. 2000. Similarities and differences. In *Proceedings of SCI2000*, pages 331–337, Orlando, FL, USA, July.

Cameron Shaw Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the 4th International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 1–12, Trento, Italy, October.

Emmanuel Giguet and Pierre-Sylvain Luquet. 2006. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 271–278, Sydney, Australia.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*, volume 1, pages 48–54, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceddings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, Pennsylvania, USA, July.

Robert Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, British Columbia, October.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, College Park, Maryland, USA.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152, Las Palmas de Gran Canaria, Spain.