Evaluating MT with Translations or Translators. What is the Difference?

Martin Volk and Søren Harder
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
volk@ling.su.se

Abstract

This paper describes a project on building a Machine Translation system for television and film subtitles. We report on the specific properties of the text genre, the language pair Swedish-Danish, and the large training corpus. We focus on the evaluation of the system output against independent and post-edited translations. We show that evaluation results against post-edited translations are higher by a margin of up to 19 points BLEU score.

1 Introduction

We are building a Machine Translation system for translating film subtitles from Swedish to Danish in a commercial setting. Most programmes are originally in English and receive Swedish subtitles based on the English video and audio (sometimes accompanied by an English manuscript). The creation of the Swedish subtitle is a manual process done by specially trained subtitlers following company-specific guidelines. In particular the subtitlers have to set the time codes (beginning and end time) for each subtitle. They use an in-house tool which allows them to attach the subtitle to specific frames in the video.

The Danish translator subsequently has access to the original English video and audio but also to the Swedish subtitles and the time codes. In most cases the translator will reuse the time codes and insert the Danish subtitle. She can, on occasion, change the time codes if she deems them inappropriate for the Danish text.

Our task is to produce draft Danish translations to speed up the translators' work. This project of automatically translating subtitles from Swedish to Danish benefits from three favourable conditions:

1. Subtitles are short textual units with little internal complexity.
2. Swedish and Danish are two closely related languages.
3. We have access to large numbers of Swedish subtitles and human-translated Danish subtitles. Their correspondence can easily be established via the time codes (which means that they are aligned on the subtitle level).

But there are also aspects of the task that are less favorable. Subtitling involves not only translation between languages, but also between modes. Subtitles are not transcriptions, but written representations of spoken language. This means firstly that the linguistic structure of the Danish and Swedish subtitles is closer to written language than the original English speech, and secondly that the original spoken content usually has to be condensed. The condensation is done by the Swedish subtitler independent of the Swedish to Danish translation, but linguistic differences between the languages (and differences between the sensibilities of the translators) may result in differences in the two languages.

The task of translating subtitles also differs from most other machine translation applications

in that we are dealing with creative language and thus we are closer to literary translation than technical translation. This is obvious in cases where rhyming song-lyrics or puns are involved, but also when the subtitler applies his individual linguistic feeling to achieve a natural and appropriate wording which blends into the video without disturbing. Finally the language of subtitling covers many domains from educational programs on any conceivable topic to exaggerated modern youth language.

We have decided to build a statistical MT (SMT) system in order to deliver a working system after a short development time and in order to best exploit the existing translations. We have trained a SMT system by using GIZA++ (Och and Ney, 2004)[1] for the alignment, Thot (Ortiz-Martínez et al., 2005)[2] for phrase-based SMT, and Phramer[3] as the decoder.

We will first present our setting and our approach for training the SMT system. We will then focus on two evaluation strategies. We first evaluated the MT output against a left-aside set of previous human translations. We computed BLEU scores (Papineni et al., 2001) of around 56 in these experiments.

In addition we computed the percentage of exactly matching subtitles against a previous human translation (How often does our system produce the exact same subtitle as the human translator?), and we compute the percentage of subtitles with a Levenshtein distance of up to 5 (meaning that the system output has an editing distance of at most 5 keystrokes from the human translation). We found that 15% of the subtitles in our evaluation corpus were translated by our system in exactly the same way (including line-breaks) as by the human translator. In addition we found that another 16% of the evaluation subtitles were 5 keystrokes or less away from the gold standard human translation.

But many automatically produced subtitles which were more than 5 keystrokes away from the human translation looked like good translations at manual inspection. Therefore we ran a set of experiments with translators who were asked to post-edit the system output. This paper describes our experience in building a large scale MT system and tells the story of the difference in evaluation results when comparing the system output to independent translations and to post-edited translator output.

2 Background

In this section we describe the parameters of our project and position it with respect to related approaches.

2.1 Characteristics of Subtitles

Foreign films and series shown in Swedish and Danish TV are usually subtitled rather than dubbed. Therefore the demand for Swedish and Danish subtitles is high. These subtitles are meant for the general public in contrast to subtitles that are specific for the hearing impaired. Those subtitles include descriptions of sounds, noises and music which we do not have to deal with.

The subtitles in our corpus are limited to 37 characters per line and usually to two lines. Depending on their length, they are shown on the screen between 2 and 8 seconds. Subtitles thus typically consist of one or two short sentences (with an average number of 10 tokens per subtitle). Sometimes a sentence spans more than one subtitle. It is then ended with a hyphen and resumed with a hyphen at the beginning of the next subtitle (this occurs about 35.7 times for each 1000 subtitles in our corpus). Example 1 shows a pair of subtitles that are close translation correspondences (although the Danish translator has decided to break the subtitle into three sentences from the two Swedish sentences).[4]

(1) SV: Det är slut, vi hade förfest här. Jätten drack upp allt.
DA: Den er væk. Vi holdt en forfest. Kæmpen drak alt.
EN: It is gone. We had a pre-party here. The giant drank it all.

In contrast, the pair in 2 exemplifies a slightly different wording chosen by the Danish translator.

[1] GIZA++ is accessible at http://www.fjoch.com/GIZA++.html

[2] Thot is available at http://thot.sourceforge.net/

[3] Phramer was written by Marian Olteanu and is available at http://www.olteanu.info/

[4] In this example and in all subsequent subtitle examples the English translations were added by the authors.

(2) SV: Där ser man vad framgång kan göra med en ung person.
DA: Der ser man, hvordan succes ødelægger et ungt menneske.
EN: There you see, what success can do to a young person / how success destroys a young person.

Typologically Swedish and Danish are very close. But of course this does not mean that the Danish subtitle translator has always chosen the same structure as the Swedish subtitler. Sometimes a Swedish subtitle consists of one sentence while its Danish counterpart consists of two sentences. This may occur either because the Danish translator has decided to split the content of the Swedish sentence into two sentences, or because he has added an extra sentence with more content, to make the text easier to understand. This extra content may represent something in the English original or not.

Similar situations occur when the Danish translator joins Swedish sentences or removes sentences that he deems unnecessary for comprehension. So even when the corresponding subtitles both consist of two sentence, this does not guarantee that sentence 1 in Swedish corresponds exactly to sentence 1 in Danish. Therefore we do not split the subtitles into sentences but we rather treat each subtitle as one textual unit.

There is one possible exception to this rule. Dialog subtitles (Swedish: *dubbelrepliker*) are candidates for splitting. Such a dialog subtitle consists of two utterances by two speakers, each in a separate line and introduced by a dash. If both the Swedish and the corresponding Danish subtitle follow this pattern, it is usually safe to treat each utterance as a separate textual unit.

(3) SV: -Min tjej vill dricka champagne.
-Titta i minibaren.
DA: - Min pige vil have champagne.
- Se i minibaren.
EN: -My girl wants to have champagne.
-Look in the minibar.

This paper can only give a rough characterization of subtitles. A more comprehensive description of the linguistic properties of subtitles can be found in (de Linde and Kay, 1999). (Gottlieb, 2001) describes the peculiarities of subtitling in Scandinavia.

2.2 Swedish and Danish in Comparison

Swedish and Danish are closely related Germanic languages. Vocabulary and grammar are similar, however orthography differs considerably, word order differs somewhat and, of course, pragmatics avoids some constructions in one language that the other language prefers. This is especially the case in the contemporary spoken language, which accounts for the bulk of subtitles.

Most Swedes and Danes understand the written language of the other. But both Danes and Swedes have to make a deliberate effort to understand each others' spoken language. Some people claim that spoken language understanding is asymmetric in that the Danes find it easier to understand Swedes than vice versa.

Let us describe some differences between Swedish and Danish that are relevant for our project. For example, there are a few word order differences. In Swedish the verb takes non-nominal complements before nominal ones, where in Danish it is the other way round. The core problem can be seen in example 4 where the verb particle immediately follows the verb in Swedish but is moved to the end of the clause in Danish.

(4) SV: Du häller ut krutet.
DA: Du hælder krudtet ud.
EN: You are pouring out the gunpowder.

A similar difference occurs in positioning the negation adverb (SV: inte, DA: ikke). It follows immediately after the verb in Swedish but appears after the object in Danish.

(5) SV: På vägen till verkstaden funkade inte oljetrycksmätaren.
DA: På vej til værkstedet fungerede olietryksmåleren ikke.
EN: On the way to the garage the oil pressure meter did not work.

In Danish there is a distinction between the use of 'der' (*there*) and 'det' (*it*) which does not exist in Swedish.

(6) SV: Alla ska vara vänner och det ska vara fred.
DA: Alle skal være venner, og der skal være fred.
EN: Everybody has to be friends and there ought to be peace.

Both Swedish and Danish mark definiteness with a suffix on nouns, but Danish does not have the double definiteness marking of Swedish. Danish nouns are not morphologically marked for definiteness when they occur with a definite article, possessive pronoun or the like.

(7) SV: Den gyllene björnen, den gyllene snyggingen.
DA: Den gyldne bjørn (/*bjørnen). Den gyldne steg (/*stegen).
EN: The golden bear, the golden hunk.

2.3 Our Subtitle Corpus

Our corpus consists of TV subtitles from soap operas (like daily hospital series), detective series, animation series, comedies, documentaries, feature films etc. In total we have access to more than 14,000 subtitle files (= single TV programmes) in each language, corresponding to about 5 million subtitles.

When we compiled our corpus we included only subtitles with matching time codes. If the Swedish and Danish time codes differed more than a threshold of 15 TV-frames (0.6 seconds) in either start or end-time, we suspected that they are not good translation equivalents and excluded them from the subtitle corpus. This idea is supported by the fact that even within the 15 frames window, the larger the time code difference between the two languages, the less confident our system gets with regards to the alignments of the subtitles.

In a first profiling step we investigated the vocabulary size of the corpus. We deleted all punctuation symbols and numbers and then counted all word form types. We found that the Swedish subtitles account for around 360,000 word form types. Interestingly, the number of Danish word form types is about 7% lower and the Danish subtitles have around 7% more tokens. We believe that this may be an artifact of the translation direction from Swedish to Danish which may lead the translator to a restrictive Danish word choice.

Another interesting profiling feature is the repetitiveness of the subtitles. We found that 28% of all Swedish subtitles in our training corpus occur more than once. Half of these multiple-occurring subtitles have exactly one Danish translation. The other half have two or more different Danish translations which are due to context differences combined with the high context dependency of short utterances and the Danish translators choosing less compact representations.

2.4 Related Projects

We have found surprisingly few other projects on automatic translation of subtitles.

(Armstrong et al., 2006) have 'ripped' subtitles (40,000 sentences) German and English as training material for their EBMT system and compared the performance to the same amount of Europarl sentences (which have more than three times as many tokens!). Training on the subtitles gave slightly better results when evaluating against subtitles, compared to training on Europarl and evaluating against subtitles. This is not surprising, although the authors point out that this contradicts some earlier findings that have shown that heterogeneous training material works better.

They do not discuss the quality of the ripped translations nor the quality of the alignments (which we found to be a major problem when we did similar experiments with English-Swedish subtitles that we had downloaded from http://www.opensubtitles.org/).

The BLEU scores are on the order of 11 to 13 for German to English (and worse for the opposite direction). Thus they are very low. They also did user evaluations with 4-point scales for intelligibility and accuracy. They asked 5 people per language pair to rate a random set of 200 sentences of system output. The judges rated English to German translations higher than the opposite direction (which contradicts the BLEU scores). But due to the small scale of the evaluation it seems premature to draw any conclusions.

(Melero et al., 2006) combined Translation Memory technology with Machine Translation, which looks interesting at first sight. But then it turns out that their Translation Memories for the language pairs Catalan-Spanish and Spanish-English were not filled with subtitles but rather

with newspaper and UN texts. They don't give any motivation for this.

The paper contains a short section on "compression" which should probably be called "(named) entity classification". The most interesting aspect in this part is that they use a parameter for translation quality threshold (for their TM lookup) and "Number of candidate translations retrieved". But disappointingly they did not train their own MT system but rather worked only with free-access web-based MT systems.

They showed that a combination of Translation Memory with such free-access web-based MT systems works better than the web-based MT systems alone. For English to Spanish this resulted in an improvement of around 7 points BLEU scores (but hardly any improvement at all for English to Czech).

It was also difficult to find other projects on Swedish to Danish Machine Translation. (Koehn, 2005) has trained his system on a parallel corpus of more than 20 million words from the European parliament. In fact he trained on all combinations of the 11 languages of the Europarl corpus. This corpus contains 27.1 million Danish tokens, but only 23.5 million Swedish tokens. The difference is due to the fact that some chapters are not translated into Swedish (there are 4,120 Danish chapters, but only 3,627 Swedish chapters).

(Koehn, 2005) reports a BLEU score of 30.3 for Swedish to Danish translation which ranks somewhere in the middle when compared to other language pairs from the Europarl corpus. The worst score was for Dutch to Finnish (10.3) and the best for Spanish to French translations (40.2). It is also noteworthy, that the scores are asymmetric. The translation direction Danish to Swedish received a BLEU score of 28.3.

3 Training the MT System

From our subtitle corpus we have set aside a random selection of files for training, consisting of 5 million subtitles, only 4 million of which are used in the current training. From the remaining part of the corpus, we have selected 24 files (approximately 10,000 subtitles) representing the diversity of the corpus from which a random selection of 1000 subtitles was taken for our test set. Before the training we prepared the data by tokenizing the text (e.g. separating punctuation symbols from words), by converting all uppercase words into lower case, and by disambiguating and slightly normalizing the use of punctuation, numbers and hyphenated words.

The amount of training data correlates with the training time. Training a system based on half a million subtitles took 16 hours on a computer with 8 GByte RAM, while training on 1 million subtitles took 3-4 days, and training on 2 million subtitles took 2-3 weeks.

We had mentioned in the introduction that we observed BLEU scores of 56 and we computed 15% exact subtitle matches plus 16% subtitles with a Levenshtein difference of 5 or less. This result was achieved by training on 4 million subtitles and evaluating against all our 1000 subtitles in the evaluation corpus. We chose the Levenshtein difference of 5 since we consider these still to be 'good' translations. Such a small difference between the system output and the human reference translation can be due to punctuation, to inflectional suffixes (e.g. the plural -s in example 8 with MT being our system output and HT the human translation) or to incorrect pronoun choices (as in example 9). Translations that only differ in the placement of line-breaks, get a Levenshtein distance of 0 (but do not count as an exact match) as line-breaks are ignored in computing the Levenshtein score.

(8) MT: Det gør ikke noget. Jeg prøver gerne hotdog med kalkun -
HT: Det gør ikke noget. Jeg prøver gerne hotdog**s** med kalkun**,** -
EN: That does not matter. I like to try hotdog(s) with turkey.

(9) MT: Jaså? Jeg kunne ikke genkende dem.
HT: Jaså? Jeg kunne ikke genkende **det**.
EN: Really? I did not recognize them / it.

A more detailed break-down of the scores revealed that there were considerable score differences between the evaluation files. The best file scored with 28.4% exact matches plus 22.8% Levenshtein 5 difference, while the worst file had 2.6% exact matches plus 6.5% Levenshtein 5 difference. It is difficult to precisely identify the reasons for these differences. One reason certainly

is the length difference in subtitles. The best file had subtitles with an average length of 10.8 tokens whereas the worst file had subtitles with 13.3 tokens on average. But also the genre and content of the file is important, with the worst file being a comedy based on puns, and the best being a drama-documentary on teaching language to a chimpanzee.

3.1 Combining Corpora

Because of the long training times we have developed a method called 'patchworking' to run the process on sub-corpora and combine these into larger models. We tested this method on 8 smaller corpora of 0.5 million blocks, and three larger corpora (2 * 1 million and 1 * 2 million), that were created by concatenating the smaller ones. This approach included a filter after running GIZA with the goal of removing 'bad' data.

This approach looked promising but in the end we found that the improvements were mainly due to two aspects that are independent of combining the corpora. Firstly, in the original setup the same corpus was used to train the translation model and the language model. Some of the later improvement was due to the larger language model based on the larger corpora. Secondly, we also observed some improvement when filtering on a 0.5 million corpus, without combining multiple corpora. So, to some extent the improvement was due to better quality of data and not larger quantity. This means that we have no clear evidence showing that using a 4 million subtitle training corpus is better than using a 0.5 million subtitle corpus.

Nevertheless we have decided to stick to the patchwork model, since we believe it is more robust, and allows for partial reuse of old models when we retrain the system. However one consequence is clear, it is advantageous to create the language model independently of the translation model.

3.2 Unknown Words

Although we have a large training corpus, there are still unknown words (= words not seen in the training data) in the evaluation data. They comprise proper names of people or products, rare word forms, compounds, and spelling deviations. Proper names need not concern us in this context since the system will copy unseen proper names (like all other unknown words) into the Danish output which in almost all cases is correct.

Rare word forms and compounds are more serious problems. It is hardly ever the case that all forms of a Swedish verb occur in our corpus (regular verbs have 7 forms accounting for active and passive in infinitive, participle, present and past tense). So even if 6 forms of a Swedish verb have been seen frequently with clear Danish translations, the 7th will be regarded as an unknown if it is missing in the training data.

Both Swedish and Danish are compounding languages which means that compounds are spelled as orthographic units and that new compounds are dynamically created. This results in unseen Swedish compounds when translating new subtitles, although often the parts of the compounds were present in the training data. We therefore generate a translation suggestion for an unseen Swedish compound by combining the Danish translations of its parts.

Variation in graphical formatting also occurs. Consider spell-outs, where spaces, commas, hyphens or even full stops are used between the letters of a word, like 'I will n o t do it', 'Jerry Seinfeld' spelled 'S, e, i, n, f, e, l , d' or 'W E L C O M E T O L A S V E G A S', or spelling variations like *ä-ä-älskar* or *abso-jävla-lut* which could be rendered in English as *lo-o-ove* or *abso-damned-lutely*. Subtitlers introduce such deviations to emphasize a word or to mimic a certain pronunciation. We handle some of these phenomena in pre-processing, but, of course, we cannot catch all of them due to their great variability.

4 Evaluating MT Performance

In the first phase of the project we evaluated the performance of our SMT system against previously translated subtitles. Many of the subtitles that did not score as good translations, because they had a Levenshtein distance greater than 5, still looked like acceptable or even equally good translations to us. Therefore we have started a new series of evaluations in which the Danish translators post-edit our system output rather than translate from scratch. We carefully picked files from different genres. And we made sure that the translators had not translated the same file before.

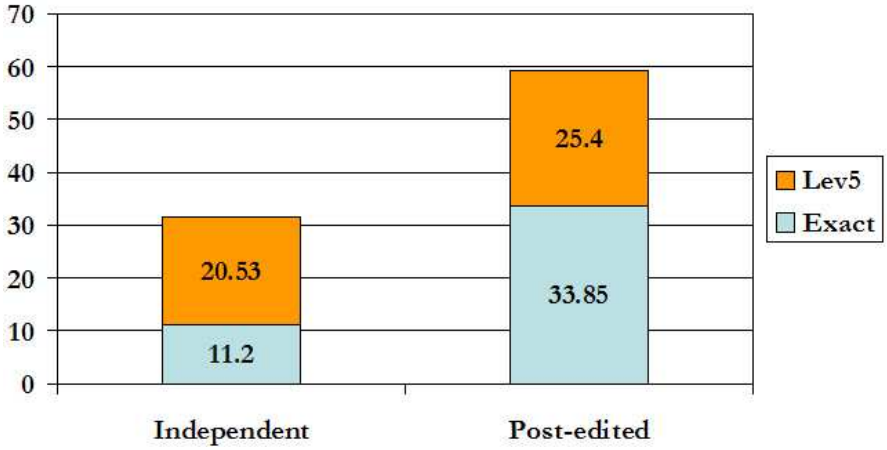

Figure 1: Results for 5 files translated independently of our system versus post-edited by translator K

We compared our system output to the post-edited files for two different translators. Translator K post-edited 5 files. For these files we computed a BLEU score of 73.5 (ranging from 67.7 to 77.7 depending on the file), and we computed 33.9% exact subtitle matches (variation from 19.1 to 45.5%) plus 25.35% subtitles within a Levenshtein distance of 5 (cf. figure 1). So we claim that our system is producing 59.25% 'good' translations compared to this translator (variation from 48% to 75.8%).

Translator L post-edited 3 files (two of which were the same files as for translator K). The resulting scores for those 3 files were lower: a BLEU score of 56.1 (ranging from 48.2 to 69.4), 16.1% exact matches (from 8.1% to 31%) with an additional 16% within a Levenshtein distance of 5, summing to 32.1% good translations (variation: 22.4 to 45.8%)

According to the translators, they found the files containing 'unscripted' language (e.g. talk-shows) and technical language[5] needed more post-editing. These observations fit with our measures, as these files scored lowest, together with comedy, which scored slightly lower than drama.

We do not have a good explanation for the large difference between the scores for the two translators. The resulting translations seem to be of similar quality. We believe that translator L may have relied more on the original English dialog than translator K. There is also a slight difference in the 'difficulty' of the files: the two overlapping files that both translators worked on were the files where translator K had his lowest and third lowest scores. So, it is likely that translator L accidentally picked the 'difficult' files from the set. But if we compare the two translators on the two files that they both post-edited, we still see that they only agree in 16.6% of the cases (31% if we include translations within a Levenshtein distance of 5). We need further studies to explain these differences between the translators.

The scores for the 5 files post-edited by K are significantly higher than our evaluation scores computed against independent translations of the same files. These 5 files received a BLEU score of 53.9 when we compared to previous independent translations, and we computed exact matches of 11.2% plus 20.5% within a Levenshtein distance of 5 (adding up to 31.7% 'good' translations). This difference is illustrated in figure 1.

The difference is not as striking for the 3 files done by translator L. When comparing the system output to previous independent translations for these 3 files we computed a BLEU score of 56.3 (which is about the same as for the post-edited version), and 8.4% exact matches plus 19.4% Levenshtein 5, adding up to 27.8% 'good' translations (which is 4.3% points lower than for the post-edited version). This confirms our hypothesis that many of the automatically translated subtitles are good translations even though they did not match the independent human translation.

[5]It is worth noting that the 'technical' language related to cars, and we are not aware that the training corpus contained any files within this domain.

Another interesting number is the inter-translator agreement (i.e. the comparison of the post-edited file with the independent translation): For translator K we computed a BLEU score of 58.4/58.7 (depending on the direction of comparison), 14% exact matches plus 21% within Levenshtein distance of 5, summing to 35% 'good' correspondences. This measure tells us how well the human translator would have done, if he had been measured with the same criteria as the Machine Translation system, and thus gives an indication of the quality of this evaluation measure. The inter-translator agreement is an upper-bound: if the 'SMT to independent translation' scores are higher than the 'inter-translator' score, then the SMT system does a better job than our translators, according to our measure. The closer the scores get to this number, the lower the quality of the measure is. We conjecture that this 'magic number' (BLEU 59) is probably lower for our corpus than for other types of corpora, as the margins of acceptable difference is probably larger in a corpus of subtitles, since it consists of creative language that has to abide to the restrictions and possibilities of conciseness and contextuality.

5 Conclusions

We are working on a Machine Translation system for translating Swedish subtitles to Danish. We have shown that evaluating the system against independent translations does not give a true picture of the translation quality and thus of the usefulness of the system. Evaluation BLEU scores were about 19 points higher when we compared our system output against post-edited translations by one of our translators. Exact matches and Levenshtein 5 scores were also clearly higher (18% higher for translator K, and 4.3% higher for translator L). These findings provide an important argument for MT developers who have only access to previous (= independent) translations. The true translation quality of your system will be clearly higher if you are working under similar conditions (in particular with large amounts of training material).

The development of the current Swedish-to-Danish SMT system has been limited by the absence of feedback-translations. Feedback-translations are subtitles that are created by a human translator correcting the system output. When there are enough feedback data available, we expect improved performance when retraining the system with these. As the random variation in the corpus diminishes, the meaningful variation will be easier to find.

We hope that the employment of our system with a large number of qualified translators / post-editors who evaluate the system output through their daily work will result in both new insights into the usability of the MT evaluation metrics, and also in a parallel corpus that is even better suitable for training a machine translation system.

References

Stephen Armstrong, Andy Way, Colm Caffrey, Marian Flanagan, Dorothy Kenny, and Minako O'Hagan. 2006. Improving the quality of automated DVD subtitles via example-based machine translation. In *Proc. of Translating and the Computer 28*, London. Aslib.

Zoe de Linde and Neil Kay. 1999. *The Semiotics of Subtitling*. St. Jerome Publishing, Manchester.

Henrik Gottlieb. 2001. Texts, translation and subtitling - in theory, and in Denmark. In Henrik Holmboe and Signe Isager, editors, *Translators and Translations*, pages 149–192. Aarhus University Press. The Danish Institute at Athens.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT-Summit*, Phuket.

Maite Melero, Antoni Oliver, and Toni Badia. 2006. Automatic multilingual subtitling in the eTITLE project. In *Proc. of Translating and the Computer 28*, London. Aslib.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. 2005. Thot: A toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*, Phuket. AAMT.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Almaden.