Martin Volk

The Automatic Translation of Film Subtitles. A Machine Translation Success Story?

1 Introduction

Every so often one hears the complaint that 50 years of research in Machine Translation (MT) has not resulted in much progress, and that current MT systems are still unsatisfactory. A closer look reveals that web-based general-purpose MT systems are used by thousands of users every day. And, on the other hand, special-purpose MT systems have been in long-standing use and work successfully in particular domains or for specific companies.

This paper[1] investigates whether the automatic translation of film subtitles can be considered a machine translation success story. We describe various projects on MT of film subtitles and contrast them to our own project in this area. We argue that the text genre "film subtitles" is well suited for MT, in particular for Statistical MT. But before we look at the translation of film subtitles let us retrace some other MT success stories.

Hutchins (1999) lists a number of successful MT systems. Amongst them is Météo, a system for translating Canadian weather reports between English and French which is probably the most quoted MT system in practical use. References to Météo usually remind us that this is a "highly constrained sublanguage system". On the other hand there are general purpose but customer-specific MT systems like the English to Spanish MT system at the Pan American Health Organization or the PaTrans system which Hutchins (1999) calls "... possibly the best known success story for custom-built MT". PaTrans was developed for LingTech A/S to translate English patents into Danish.

Earlier (Whitelock and Kilby (1995), p.198) had called the METAL system "a success story in the development of MT". METAL is mentioned as "successfully used at a number of European companies" (by that time this meant a few dozen installations in industry, trade or banking). During the same time the European Union has been successfully using a customized version of Systran for its translation service but also later for online access by all its employees. Broad coverage systems like METAL and Systran have always resulted in a translation quality that required post-editing before publications.

[1] This is a slightly corrected and updated version of a paper first published in: Joakim Nivre, Mats Dahllöf, Beáta Megyesi (Eds.) *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, Uppsala University, 2008.

Attempts to curb the post-editing by pre-editing or constraining the source documents have gone under the name of controlled language MT. Hutchins (1999) mentions controlled language MT (e.g. at the Caterpillar company) as an example of successful employment of MT. This is an area where part of the pioneering work was done at Uppsala University by Anna Sågvall Hein and her group (Almqvist and Sågvall Hein, 1996), including the development of controlled Swedish for the automobile industry. This research subsequently led to a competitive MT system for translating from Swedish to English (Sågvall Hein et al., 2002).

The claim that web-based machine translation is a success is based on the fact that it is used by large numbers of users. Critics do not subscribe to this argument as long as the translation quality is questionable. Still, popular services including Systran (www.systran.co.uk with 14 source languages) and Google (www.google.com/translate_t with 21 language pairs) cover major Western languages like English, Spanish and French, but also Arabic and Chinese. On the other hand there are providers that have successfully occupied niche language pairs like Danish to English (Bick, 2007).

So we see that MT success stories vary considerably. We regard the following criteria as the main indicators of success:

1. A large user base (this criterion is used in web-based MT services for the general public)
2. Customer satisfaction (this criterion is used in customer-specific MT systems and usually based on improved productivity and return on investment)
3. Long-term usage of the MT system

We will check which of these criteria apply to the automatic translation of film subtitles.

2 Characteristics of Film Subtitles

When films are shown to audiences in language environments that differ from the language spoken in the film, then some form of translation is required. Larger markets like Germany and France typically use dubbing of foreign films so that it seems that the actors are speaking the local language. Smaller countries often use subtitles. Pedersen (2007) discusses the advantages and drawbacks of both methods.

Foreign films and series shown in Scandinavian TV are usually subtitled rather than dubbed. Therefore the demand for Swedish, Danish, Norwegian and Finnish subtitles is high. These subtitles are meant for the general public in contrast to subtitles that are specific for the hearing-impaired which often include descriptions of sounds, noises and music. Subtitles also differ with respect to whether they are produced online (e.g. in live

talkshows or sport reports) or offline (e.g. for pre-produced series). This paper focuses on general public subtitles that are produced offline.

In our machine translation project, we use a parallel corpus of Swedish, Danish and Norwegian subtitles. The subtitles in this corpus are limited to 37 characters per line and usually to two lines.[2] Depending on their length, they are shown on screen between 2 and 8 seconds. Subtitles typically consist of one or two short sentences with an average number of 10 tokens per subtitle in our corpus. 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 human-translated pair of subtitles that are close translation correspondences although the Danish translator has decided to break the two sentences of the Swedish subtitle into three sentences.[3]

(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.

(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.*

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) and (Díaz-Cintas and Remael, 2007). Gottlieb (2001) and Pedersen (2007) describe the peculiarities of subtitling in Scandinavia.

3 Approaches to the Automatic Translation of Film Subtitles

In this section we describe other projects on the automatic translation of subtitles. We distinguish between rule-based, example-based, and statistical approaches.

[2] Although we are working on both Swedish to Danish and Swedish to Norwegian MT of subtitles, this paper focuses on translation from Swedish to Danish. The issues for Swedish to Norwegian are the same to a large extent.

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

3.1 Rule-based MT of Film Subtitles

Popowich et al. (2000) provide a detailed account of a MT system tailored towards the translation of English subtitles into Spanish. Their approach is based on a MT paradigm which relies heavily on lexical resources but is otherwise similar to the transfer-based approach. A unification-based parser analyzes the input sentence (including proper-name recognition), followed by the lexical transfer which provides the input for the generation process in the target language (including word selection and correct inflection).

Popowich et al. (2000) mention that the subtitle domain has certain advantages for MT. According to them it is advantageous that output subtitles can and should be grammatical even if the input sometimes is not. They argue that subtitle readers have only a limited time to perceive and understand a given subtitle and that therefore grammatical output is essential. And they follow the strategy that "it is preferable to drop elements from the output instead of translating them incorrectly" (p.331). This is debateable and opens the door for incomplete output.

Although Popowich et al. (2000) call their system "a hybrid of both statistical and symbolic approaches" (p.333), it is a symbolic system by today's standards. The statistics are only used for efficiency improvements but are not at the core of the methodology. The paper was published before automatic evaluation methods were invented. Instead Popowich et al. (2000) used the classical evaluation method where native speakers were asked to judge the grammaticality and fidelity of the system. These experiments resulted in "70% of the translations ... be ranked as correct or acceptable, with 41% being correct" which is an impressive result. Whether this project can be regarded as a MT success story depends on whether the system was actually employed in production. This information is not provided in the paper.

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 articles and UN texts. They don't give any motivation for this. And disappointingly they did not train their own MT system but rather worked only with free-access web-based MT systems (which we assume are rule-based systems).

They showed that a combination of Translation Memory with such web-based MT systems works better than the web-based MT systems alone. For English to Spanish translation this resulted in an improvement of around 7 points in BLEU scores (Papineni et al., 2001) but hardly any improvement at all for English to Czech.

3.2 Example-based MT of Film Subtitles

Armstrong et al. (2006) "ripped" subtitles (40,000 sentences) German and English as

training material for their Example-based MT 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 freely available English-Swedish subtitles).

The BLEU scores are on the order of 11 to 13 for German to English (and worse for the opposite direction). These are very low scores. They also conducted 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). Owing to the small scale of the evaluation, however, it seems premature to draw any conclusions.

3.3 Statistical MT of Film Subtitles

Descriptions of Statistical MT systems for subtitles are practically non-existent, probably due to the lack of freely available training corpora. Until recently there were no freely available subtitle collections. Both Tiedemann (2007) and Lavecchia et al. (2007) report on efforts to build such corpora with alignment on the subtitles.

Tiedemann (2007) works with a huge collection of subtitle files that are available on the internet at www.opensubtitles.org. These subtitles have been produced by volunteers in a great variety of languages. But the volunteer effort also results in subtitles of often dubious quality (they include timing, formatting, and linguistic errors). The hope is that the enormous size of the corpus will supersede the noise in practical applications. The first step then is to align the files across languages on the subtitle level. The time codes alone are not sufficient as different (amateur) subtitlers have worked with different time offsets and sometimes even different versions of the same film. Still, Tiedemann (2007) shows that an alignment approach based on time overlap combined with cognate recognition is clearly superior to pure length-based alignment. He has evaluated his approach on English, German and Dutch. His results of 82.5% correct alignments for Dutch-English and 78.1% correct alignments for Dutch-German show how difficult the alignment task is. And a rate of around 20% incorrect alignments will certainly be problematic when training a Statistical MT system on these data.

Lavecchia et al. (2007) also work with subtitles obtained from the internet. They work on French-English subtitles and use a method which they call Dynamic Time Warping for aligning the files across the languages. This method requires access to a bilingual

dictionary to compute subtitle correspondences. They compiled a small test corpus consisting of 40 subtitle files, randomly selecting around 1300 subtitles from these files for manual inspection. Their evaluation focused on precision while sacrificing recall. They report on 94% correct alignments when turning recall down to 66%. They then go on to use the aligned corpus to extract a bilingual dictionary and to integrate this dictionary in a Statistical MT system. They claim that this improves the MT system with 2 points BLEU score (though it is not clear which corpus they have used for evaluating the MT system).

This summary indicates that most work on the automatic translation of film subtitles with Statistical MT is still in its infancy. Our own efforts are larger and have resulted in a mature MT system. We will report on them in the following section.

4 The Stockholm MT System for Film Subtitles

We have built Machine Translation systems for translating film subtitles from Swedish to Danish (and Swedish to Norwegian) in a commercial setting. Some of this work has been described earlier by Volk and Harder (2007).

Most films 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 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 or Norwegian 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 target language subtitle. She can, on occasion, change the time codes if she deems them inappropriate for the target language.

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

1. Subtitles are short textual units with little internal complexity (as described in section 2).

2. Swedish, Danish and Norwegian are closely related languages.

3. We have access to large numbers of Swedish subtitles and human-translated Danish and Norwegian subtitles. Their correspondence can easily be established via the time codes which leads to an alignment on the subtitle level.

But there are also aspects of the task that are less favorable. Subtitles are not transcriptions, but written representations of spoken language. As a result the linguistic structure of subtitles is closer to written language than the original (English) speech, and the original spoken content usually has to be condensed by the Swedish subtitler.

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 linguistic intuitions to achieve a natural and appropriate wording which blends into the video without disturbing. Finally, the language of subtitling covers a broad variety of 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 shorten the development time (compared to a rule-based system) and in order to best exploit the existing translations. We have trained our SMT system by using GIZA++ (Och and Ney, 2004)[4] for the alignment, Thot (Ortiz-Martínez et al., 2005)[5] for phrase-based SMT, and Phramer[6] as the decoder.

We will first present our setting and our approach for training the SMT system and then describe the evaluation results.

4.1 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.

One of the relevant differences for our project concerns word order. 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 3 where the verb particle *ut* immediately follows the verb in Swedish but is moved to the end of the clause in Danish.

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

A similar word order difference occurs in positioning the negation adverb (SV: *inte*, DA: *ikke*). Furthermore, Danish distinguishes between the use of *der* (EN: *there*) and *det*

[4]GIZA++ is accessible at http://www.fjoch.com/ GIZA++.html
[5]Thot is available at http://thot.sourceforge.net/
[6]Phramer was written by Marian Olteanu and is available at http://www.olteanu.info/

(EN: *it*) but Swedish does not. Both Swedish and Danish mark definiteness with a suffix on nouns, but Danish does not have the double definiteness marking of Swedish.

4.2 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 more than 5 million subtitles (equalling more than 50 million words).

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 were not good translation equivalents and excluded them from the subtitle corpus. In this way we were able to avoid complicated alignment techniques. Most of the resulting subtitle pairs are high-quality translations of one another thanks to the controlled workflow in the commercial setting.

In a first profiling step we investigated the vocabulary size of the corpus. After removing all punctuation symbols and numbers we counted all word form types. We found that the Swedish subtitles amounted to around 360,000 word form types. Interestingly, the number of Danish word form types is about 5.5% lower, although the Danish subtitles have around 1.5% more tokens. We believe that this difference 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 recurring 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.

From our subtitle corpus we chose a random selection of files for training the translation model and the language model. We currently use 4 million subtitles for training. From the remaining part of the corpus, we 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 tokenized the subtitles (e.g. separating punctuation symbols from words), converting all uppercase words into lower case, and normalizing punctuation symbols, numbers and hyphenated words.

4.3 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, spelling deviations and foreign words. 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. Hardly ever do all forms of a Swedish verb occur in our training corpus (regular verbs have 7 forms). 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 poses problems. 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", "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.

Foreign words are a problem when they are homographic with words in the source language Swedish (e.g. when the English word *semester* = "university term" interferes with the Swedish word *semester* which means "vacation"). Example 4 shows how different languages (here Swedish and English) are sometimes intertwined in subtitles.

(4) SV: Hon gick ut Boston University's School of the Performing Arts-
-och hon fick en dubbelroll som halvsystrarna in "As the World Turns".
EN: *She left Boston University's School of the Performing Arts and she got a double role as half sisters in "As the World Turns".*

4.4 Evaluating the Performance of the Stockholm MT System

We first evaluated the MT output against a left-aside set of previous human translations. We computed BLEU scores of around 57 in these experiments. In addition we computed the percentage of exactly matching subtitles against a previous human translation (How

	Exact matches	Levenshtein-5 matches	BLEU
Crime series	15.0%	35.3%	63.9
Comedy series	9.1%	30.6%	54.4
Car documentary	3.2%	22.8%	53.6
Average	9.1%	21.6%	57.3

Table 1: Evaluation Results against a Prior Human Translation

often does our system produce the exact same subtitle as the human translator?), and we computed the percentage of subtitles with a Levenshtein distance of up to 5 which means that the system output has an editing distance of at most 5 basic character operations (deletions, insertions, substitutions) from the human translation.

We decided to use a Levenshtein distance of 5 as a threshold value as we consider translations at this edit distance from the reference text 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 5 with MT being our Danish system output and HT the human translation) or to incorrect pronoun choices.

(5) 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.*

Table 1 shows the results for three files (selected from different genres), for which we have prior translations (done independently of our system). We observe between 3.2% and 15% exactly matching subtitles, and between 22.8% and 35.3% subtitles with a Levenshtein distance of up to 5. Note that the percentage of Levenshtein matches includes the exact matches (which correspond to a Levenshtein distance of 0).

On manual inspection, however, many automatically produced subtitles which were more than 5 keystrokes away from the human translations still looked like good translations. Therefore we conducted another series of evaluations with translators who were asked to post-edit the system output rather than to translate from scratch. We made sure that the translators had not translated the same file before.

Table 2 shows the results for the same three files for which we have one prior translation. We gave our system output to six translators and obtained six post-edited versions. Some translators were more generous than others, and therefore we averaged their scores. When using post-editing, the evaluation figures are 13.2 percentage points

	Exact matches	Levenshtein-5 matches	BLEU
Crime series	27.7%	47.6%	69.9
Comedy series	26.0%	45.7%	67.7
Car documentary	13.2%	35.9%	59.8
Average	22.3%	43.1%	65.8

Table 2: Evaluation Results averaged over 6 Post-editors

higher for exact matches and 19.5 percentage points higher for Levenshtein-5 matches. It becomes also clear that the translation quality varies considerably across film genres. The crime series file scored consistently higher than the comedy file which in turn was clearly better than the car documentary.

There are only few other projects on Swedish to Danish Machine Translation (and we have not found a single one on Swedish to Norwegian). Koehn (2005) 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 in the Europarl corpus. 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). The fact that our BLEU scores are much higher even when we evaluate against prior translations (cf. the average of 57.3 in table 1) is probably due to the fact that subtitles are shorter than Europarl sentences and perhaps also due to our larger training corpus.

5 Conclusions

We have sketched the text genre characteristics of film subtitles and shown that Statistical MT of subtitles leads to good quality when the input is a large high-quality parallel corpus. We are working on Machine Translation systems for translating Swedish film subtitles to Danish and Norwegian with very good results (in fact the results for Swedish to Norwegian are slightly better than for Swedish 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 8.5 points higher when we compared our system output against post-edited translations averaged over six translators. Exact matches and Levenshtein 5 scores were also clearly higher.

We are dealing with customer-specific MT systems covering a broad set of textual domains. The customer is satisfied and has employed our MT systems in large scale subtitle production since early 2008. The MT systems have resulted in considerable time savings in the translation process. It is by now safe to call this a Machine Translation success story.

6 Acknowledgements

We would like to thank Jörgen Aasa, Søren Harder and Christian Hardmeier for sharing their expertise, providing evaluation figures and commenting on an earlier version of the paper.

References

Almqvist, I. and Sågvall Hein, A. (1996). Defining ScaniaSwedish - a controlled language for truck maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications*, Katholieke Universiteit Leuven.

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., and O'Hagan, M. (2006). Improving the quality of automated DVD subtitles via example-based machine translation. In *Proc. of Translating and the Computer 28*, London. Aslib.

Bick, E. (2007). Dan2eng: Wide-coverage Danish-English machine translation. In *Proc. of Machine Translation Summit XI*, Copenhagen.

de Linde, Z. and Kay, N. (1999). *The Semiotics of Subtitling*. St. Jerome Publishing, Manchester.

Díaz-Cintas, J. and Remael, A. (2007). *Audiovisual Translation: Subtitling*, volume 11 of *Translation Practices Explained*. St. Jerome Publishing, Manchester.

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

Hutchins, J. (1999). The development and use of machine translation systems and computer-based translation tools. In *Proc. of International Symposium on Machine Translation and Computer Language Information Processing*, Beijing.

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

Lavecchia, C., Smaili, K., and Langlois, D. (2007). Machine translation of movie subtitles. In *Proc. of Translating and the Computer 29*, London. Aslib.

Melero, M., Oliver, A., and Badia, T. (2006). Automatic multilingual subtitling in the eTITLE project. In *Proc. of Translating and the Computer 28*, London. Aslib.

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

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2005). Thot: A toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*, Phuket. AAMT.

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

Pedersen, J. (2007). *Scandinavian Subtitles. A Comparative Study of Subtitling Norms in Sweden and Denmark with a Focus on Extralinguistic Cultural References*. PhD thesis, Stockholm University. Department of English.

Popowich, F., McFetridge, P., Turcato, D., and Toole, J. (2000). Machine translation of closed captions. *Machine Translation*, 15:311–341.

Sågvall Hein, A., Forsbom, E., Tiedemann, J., Weijnitz, P., Almqvist, I., Olsson, L.-J., and Thaning, S. (2002). Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of LREC 2002. Third International Conference on Language Resources and Evaluation*, pages 1759 – 1766, Las Palmas.

Tiedemann, J. (2007). Improved sentence alignment for movie subtitles. In *Proceedings of RANLP*, Borovets, Bulgaria.

Volk, M. and Harder, S. (2007). Evaluating MT with translations or translators. What is the difference? In *Machine Translation Summit XI Proceedings*, Copenhagen.

Whitelock, P. and Kilby, K. (1995). *Linguistic and Computational Techniques in Machine Translation System Design*. Studies in Computational Linguistics. UCL Press, London, 2 edition.