# Machine Translation of TV Subtitles for Large Scale Production

**Martin Volk, Rico Sennrich**
University of Zürich
Computational Linguistics
CH-8050 Zurich
(volk|sennrich)@cl.uzh.ch

**Christian Hardmeier**
Fondazione Bruno Kessler
Human Language Technologies
I-38123 Trento
ch@rax.ch

**Frida Tidström**
University of Stockholm
Datorlingvistik
SE-10691 Stockholm
fridatidstrom@hotmail.com

## Abstract

This paper describes our work on building and employing Statistical Machine Translation systems for TV subtitles in Scandinavia. We have built translation systems for Danish, English, Norwegian and Swedish. They are used in daily subtitle production and translate large volumes. As an example we report on our evaluation results for three TV genres. We discuss our lessons learned in the system development process which shed interesting light on the practical use of Machine Translation technology.

## 1 Introduction

Media traditions distinguish between subtitling and dubbing countries. Subtitling countries broadcast TV programs with the spoken word in the original language and subtitles in the local language. Dubbing countries (like Germany, France and Spain) broadcast with audio in the local language. Scandinavia is a subtitling area and thus large amounts of TV subtitles are needed in Swedish, Danish and Norwegian.

Ideally subtitles are created for each language independently, but for efficiency reasons they are often translated from one source language to one or more target languages. To support the efficient translation we have teamed up with a Scandinavian subtitling company to build Machine Translation (MT) systems. The systems are in practical use today and used extensively. Because of the established language sequence in the company we have built translation systems from Swedish to Danish and to Norwegian. After the successful deployment of these two systems, we have started working on other language pairs including English, German and Swedish. The examples in this paper are taken from our work on Swedish to Danish. The issues for Swedish to Norwegian translation are the same to a large extent.

In this paper we describe the peculiarities of subtitles and their implications for MT. We argue that the text genre "TV subtitles" is well suited for MT, in particular for Statistical MT (SMT). We first introduce a few other MT projects for subtitles and will then present our own. We worked with large corpora of high-quality human translated subtitles as input to SMT training. Finally we will report on our experiences in the process of building and deploying the systems at the subtitling company. We will show some of the needs and expectations of commercial users that deviate from the research perspective.

## 2 Characteristics of TV Subtitles

When films, series, documentaries etc. are shown in language environments that differ from the language spoken in the video, then some form of translation is required. Larger markets like Germany and France typically use dubbing of foreign media 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.

In Scandinavian TV, foreign programs 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

Ventsislav Zhechev (ed.): *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*, pp. 53–62. Denver, CO, 4 November 2010.

descriptions of sounds, noises and music (cf. (Matamala and Orero, 2010)). 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 to two lines. 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. The first subtitle is then ended with a hyphen and the sentence is resumed with a hyphen at the beginning of the next subtitle. This occurs about 36 times for each 1000 subtitles in our corpus. TV subtitles contain a lot of dialogue. One subtitle often consists of two lines (each starting with a dash) with the first being a question and the second being the answer.

Although Swedish and Danish are closely related languages, translated subtitles might differ in many respects. 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.[1]

(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 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.*

[1] In all subtitle examples the English translations were added by the authors.

The space limitations on the screen result in special linguistic properties. For example, when we investigated English subtitles we have noticed that apostrophe-s-contractions (for "is, has, us") are particularly frequent in subtitles because of their closeness to spoken language. Examples are "He's watching me; He's lost his watch; Let's go". In a random selection of English subtitles we found that 15% contained apostrophe-s. These contractions need to be disambiguated, otherwise we end up with translations like "Oh my gosh, Nicole's dad is the coolest" being rendered in German as "Mein Gott, Nicole ist Papa ist der coolste" where the possessive 's' is erroneously translated as a copula verb. We have built a special PoS tagger for preprocessing the subtitles, which solves this problem well.

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, Nagel et al. (2009) in other European countries.

## 3 Approaches to the Automatic Translation of Film Subtitles

In this section we describe other projects on the automatic translation of subtitles.[2] We assume subtitles in one language as input and aim at producing an automatic translation of these subtitles into another language. In this paper we do not deal with the conversion of the film transcript into subtitles which requires shortening the original dialogue (cf. (Prokopidis et al., 2008)). We distinguish between rule-based, example-based, and statistical approaches.

### 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

[2] Throughout this paper we focus on TV subtitles, but in this section we deliberately use the term "film subtitles" in a general sense covering both TV and movie subtitles.

input sentence (including proper-name recognition), followed by lexical transfer which provides the input for the generation process in the target language (including word selection and correct inflection).

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. 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 ... ranked as correct or acceptable, with 41% being correct" which is an impressive result. This project resulted in a practical real-time translation system and was meant to be sold by TCC Communications as "a consumer product that people would have in their homes, much like a VCR." But unfortunately the company went out of business before the product reached the market.[3]

Melero et al. (2006) combined Translation Memory technology with Machine Translation for the language pairs Catalan-Spanish and Spanish-English but their Translation Memories were not filled with subtitles but rather with newspaper articles and UN texts. They don't give any motivation for this. 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 (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" German and English subtitles (40,000 sentences) as training material for their Example-based MT system and compared the performance to a system trained on 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). Their BLEU scores are on the order of 11 to 13 for German to English (and worse for the opposite direction).

### 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 (i.e. collections of human-translated subtitles). Both Tiedemann (2007) and Lavecchia et al. (2007) report on efforts to build such corpora with aligned 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. However the volunteer effort also results in subtitles of often dubious quality. Subtitles contain timing, formatting, and linguistic errors. The hope is that the enormous size of the corpus will still result in useful applications. The first step then is to align the files across languages on the subtitle level. 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.

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

[3] Personal communication with Fred Popowich in August 2010.

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 work on the automatic translation of film subtitles with Statistical MT is limited because of the lack of freely available high-quality training data. Our own efforts are based on large proprietary subtitle data and have resulted in mature MT systems. We will report on them in the following section.

## 4 Our MT Systems for TV Subtitles

We have built Machine Translation systems for translating film subtitles from Swedish to Danish and to Norwegian in a commercial setting. Some of this work has been described earlier by Volk and Harder (2007) and Volk (2008).

Most films are originally in English and receive Swedish subtitles based on the English video and audio (sometimes accompanied by an English transcript). 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 link 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.

We have built systems that produce Danish and Norwegian draft translations to speed up the translators' work. This project of automatically translating subtitles from Swedish to Danish and Norwegian benefited 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. The grammars are similar, however orthography differs considerably, word order differs somewhat and, of course, one language avoids some constructions that the other language prefers.

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.

There are other 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 standing out. 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 statistical MT (SMT) systems 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 systems by using standard open source SMT software. Since Moses was not yet available at the starting time or our project, we trained our systems by using GIZA++ (Och and

Ney, 2004) for the alignment, Thot (Ortiz-Martínez et al., 2005) for phrase-based SMT, and Phramer (www.olteanu.info) as the decoder.

We will first present our setting and the evaluation results and then discuss the lessons learned from deploying the systems in the subtitling company.

### 4.1 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 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 thanks to the controlled workflow in the commercial setting. Note that we are not aligning sentences. We work with aligned subtitles which can consist of one or two or three short sentences. Sometimes a subtitle holds only the first part of a sentence which is finished in the following subtitle.

In a first profiling step we investigated 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 step 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.

### 4.2 Unknown Words

Although we have a large training corpus, there are still unknown 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 target language 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. For an unseen word that is longer than 8 characters we split it into two parts in all possible ways. If the two parts are in our corpus, we gather the most frequent Danish translation of each for the generation of the target language compound. This has resulted in a measurable improvement in the translation quality. To keep things simple we disregard splitting compounds into three or more parts. These cases are extremely rare in subtitles.

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 3 shows how different languages (here Swedish and English) are sometimes intertwined in subtitles.

(3) SV: Hon gick ut Boston University's School of the Performing Arts-
-och hon fick en dubbelroll som halvsystrarna i "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".*

### 4.3 Evaluating the MT Performance

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. But BLEU scores are not very informative at this level of performance. Nor are they clear indicators of translation quality for non-technical people. The main criterion for determining the usefulness of MT for the company is the potential time-saving. Hence, we needed a measure that better indicates the post-editing effort to help the management in its decision.

Therefore 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 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 4 with MT being our Danish system output and HT the human translation) or to incorrect pronoun choices.

(4) 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 (created 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 the company's 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 higher for exact matches and 13.5 percentage points higher for Levenshtein-5 matches. It is also clearly visible 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

| | 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% | 29.6% | 58.5 |

Table 1: Evaluation Results against a Prior Human Translation

| | 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

Swedish to Danish translation which ranks somewhere in the middle when compared to other language pairs from the Europarl corpus. Newer numbers from 2008 experiments in the EuroMatrix project based on a larger Europarl training corpus (40 million words) report on 32.9 BLEU points (see http://matrix.statmt.org/matrix). Training and testing on the legislative texts of the EU (the Acquis Communautaire corpus) resulted in 46.6 BLEU points for Swedish to Danish translation. This shows that the scores are highly text-genre dependent. 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 and grammatically simpler than Europarl and Acquis sentences.

### 4.4 Linguistic Information in SMT for Subtitles

The results reported in tables 1 and 2 are based on a purely statistical MT system. No linguistic knowledge was included. We wondered whether linguistic features such as Part-of-Speech tags or number information (singular vs. plural) could improve our system. We therefore ran a series of experiments to check this hypothesis using factored SMT for Swedish - Danish translation. Hardmeier and Volk (2009) describe these experiments in detail. Here we summarize the main findings.

When we used a large training corpus of around 900,000 subtitles or 10 million tokens per language, the gains from adding linguistic information were generally small. Minor improvements were observed when using additional language models operating on part-of-speech tags and tags from morphological analysis. A technique called analytical translation, which enables the SMT system to back off to separate translation of lemmas and morphological tags (provided by Eckhard Bick's tools) when the main phrase table does not provide a satisfactory translation, resulted in slightly improved vocabulary coverage.

The results were different when the training corpus is small. In a series of experiments with a corpus size of only 9,000 subtitles or 100,000 tokens per language, different manners of integrating linguistic information were consistently found to be beneficial, even though the improvements were small. When the corpus is not large enough to afford reliable parameter estimates for the statistical models, adding abstract data with richer statistics stands to improve the behavior of the system.

The most encouraging findings were made in experiments in an asymmetric setting, where a small source language corpus (9,000 subtitles) was combined with a much larger target language corpus (900,000 subtitles). A considerable improvement to the score was realized just by adding a language model trained on the larger corpus without any linguistic annotation.

In all of our SMT work we have lumped all training data together, although we are aware that we are

dealing with different textual domains. As we have seen, the translation results for the crime series were clearly different from the translation results of the car documentary. As more human-translated subtitles come in over time, it might be advantageous to build separate MT systems for different subtitle domains.

## 5 Lessons for SMT in Subtitle Production

We have built MT systems for subtitles covering a broad range of textual domains. The subtitle company is satisfied and has been using our MT systems in large scale subtitle production since early 2008. In this section we summarize our experiences in bringing the MT systems to the user, i.e. the subtitler in the subtitling company. The subtitlers do not interact with the MT systems directly. Client managers function as liaison between the TV channels and the freelance subtitlers. They provide the subtitlers with the video, the original subtitle (e.g. in Swedish) and the draft subtitles produced by our MT systems (e.g. draft Danish subtitles). The subtitlers work as MT post-editors and return the corrected target-language subtitle file to the client manager.

**Combination of Translation Memory and SMT** From the start of the project we had planned to combine translation memory functionality with SMT. When our system translates a subtitle, it first checks whether the same subtitle is in the database of already translated subtitles. If the subtitle is found with one translation, then this translation is chosen and MT is skipped for this subtitle. If, on the other hand, the subtitle is found with multiple translation alternatives, then the most frequent translation is chosen. In case of translation alternatives with the same frequency, we randomly pick one of them.

To our surprise this translation memory lookup contributes almost nothing to the translation quality of the system. The difference is less than one percentage point in Levenshtein-5 matches. This is probably due to the fact that repetitive subtitles are mostly short subtitles of 5 words or less. Since our SMT system works with 5-grams, it will contain these chunks in its phrase table and produce a good translation. Considering the effort of setting up the translation memory database, we are unsure whether the TM-MT combination is worth the investment in this particular context.

**System Evaluation** As researchers we are interested in computing translation quality scores in order to measure progress in system development. The subtitling company, however, is mainly interested in the time savings that will result from the deployment of the translation system. We therefore measured the system quality not only in BLEU scores but also in exact matches and Levenshtein-5 distance between MT output and reference translations. These latter measures are much easier to interpret. In addition, our evaluations with six post-editors gave a clearer picture of the MT quality than comparing against a previous human translation. Still the problem persists as to what time saving the evaluation scores indicate. The post-editors themselves have given rather cautious estimates of time savings, since they are aware that in the long run MT means they will receive less money for working on a certain amount of subtitles. It is therefore important that the company creates a win-win situation where MT enables post-editors to earn more per working hour and the company still makes a higher profit on the subtitles than before.

**Integration of SMT into the Subtitling Workflow** It is of utmost importance to organize a smooth integration of the MT system into the subtitling workflow. In our case this meant that client managers will put the input file in a certain folder on the translation server and take the draft translation from another folder a few minutes later. In order to avoid duplicate work, each Swedish file is automatically translated to both Danish and Norwegian even if one of the translations is not immediately needed. The output file must be a well-formed time-coded subtitle file where no subtitle exceeds the character limit. Furthermore each long subtitle in the MT output needs to have a line break set at a "natural" position avoiding split linguistic units.

**MT Influence on Linguistic Intuition** Subtitle post-editors feared that MT output influences their linguistic intuitions. This is not likely to happen with clearly incorrect translations, but it may happen with slightly strange constructions. When a post-editor encounters such a strange wording for the first

time, she will correct it. But when the strange wording occurs repeatedly, it will not look strange any longer. The problem of source language influence has been known to translators for a long time, but it is more severe with MT output. The post-editors have to consider and edit constructions which they would never produce themselves.

We had therefore asked post-editors to report such observations to the development team, but we have not received any complaints about this. This could mean that this phenomenon is rare, or it is so subconscious that post-editors do not notice it. Targeted research is needed to investigate the long-term impact of MT output on the subtitles' linguistic characteristics.

**System Maintenance and Updates** A complex SMT system requires a knowledgable maintenance person. Maintenance comprises general issues such as restarting the system after server outages, but it also comprises fixes in the phrase table after translators complained about rude language in some translations. The systems will also profit from regular retraining as new translations (i.e. post-edited subtitles) come in. Interestingly the company is reluctant to invest man power into retraining as long as the systems work as reliably as they do. They follow the credo "never change a working system". Of course, one would also need to evaluate the new version and prove that it indeed produces better translations than the previous version. So, retraining requires a substantial investment.

**Presenting Alternative Translations** For a while we pondered whether we should present both the translation memory hit and the MT output or alternatively the three best SMT candidates to the post-editor. But post-editors distinctly rejected this idea. They have a lot of information on the screen already (video, time codes, source language subtitle). They do not want to go through alternative translation suggestions. This takes too much time.

**Suppressing Bad Translations** An issue that has followed us throughout the project is the suppression of (presumably) bad translations. While good machine translations considerably increase the productivity of the post-editors, editing bad translations is tedious and frequently slower than translating from scratch. To take away some of this burden from the post-editors, we experimented with a Machine Learning component to predict confidence scores for the individual subtitles output by our Machine Translation systems. Closely following the work by (Specia et al., 2009), we prepared a data set of 4,000 machine-translated subtitles, manually annotated for translation quality on a 1-4 scale by the post-editors. We extracted around 70 features based on the MT input and output, their similarity and the similarity between the input and the MT training data. Then we trained a Partial Least Squares regressor to predict quality scores for unseen subtitles.

Like (Specia et al., 2009), we used Inductive Confidence Machines to calibrate the acceptance threshold of our translation quality filter. We found that a confidence filter with the features proposed by Specia et al. performs markedly worse on our subtitle corpus than on the data used by the original authors. This may partly be due to the shortness of the subtitles: Since an average subtitle is only about 10 tokens long, it may be more difficult to judge its quality with text surface features than in a text with longer sentences, where there are more opportunities for matches or mismatches, so the features are more informative. Currently, we are exploring other features and other Machine Learning techniques since we are convinced that filtering out bad translations is important to increase the efficiency of the post-editors.

## 6 Conclusions

We have sketched the text genre characteristics of TV subtitles and shown that Statistical MT of subtitles leads to production strength translations when the input is a large high-quality parallel corpus. We have built Machine Translation systems for translating Swedish TV 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 systems against independent translations does not give a true picture of the translation quality and thus of the usefulness of the systems. Evaluation BLEU scores were about 7.3 points higher when we compared our MT output against post-edited translations averaged

over six translators. Exact matches and Levenshtein-5 scores were also clearly higher. First results for English to Swedish SMT of subtitles show also good results albeit somewhat lower evaluation scores.

We have listed our experiences in building and deploying the SMT systems in a subtitle production workflow. We are convinced that many of these issues are equally valid in other production environments. We plan to develop more MT systems for more language pairs as the demand for subtitles continues to increase.

## Acknowledgements

We would like to thank Jörgen Aasa and Søren Harder for sharing their expertise and providing evaluation figures.

## References

Armstrong, Stephen, 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 *Proceedings of Translating and the Computer 28*, London. Aslib.

Díaz-Cintas, Jorge and Aline Remael. 2007. *Audiovisual Translation: Subtitling*, volume 11 of *Translation Practices Explained*. St. Jerome Publishing, Manchester.

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

Hardmeier, Christian and Martin Volk. 2009. Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles. In Jokinen, Kristiina and Eckhard Bick, editors, *Proceedings of the 17th Nordic Conference of Computational Linguistics. NODALIDA*, pages 57–64, Odense, May.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, Phuket.

Lavecchia, Caroline, Kamel Smaili, and David Langlois. 2007. Machine Translation of Movie Subtitles. In *Proceedings of Translating and the Computer 29*, London. Aslib.

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

Matamala, Anna and Pilar Orero, editors. 2010. *Listening to Subtitles. Subtitles for the Deaf and Hard of Hearing*. Peter Lang Verlag.

Melero, Maite, Antoni Oliver, and Toni Badia. 2006. Automatic Multilingual Subtitling in the eTITLE Project. In *Proceedings of Translating and the Computer 28*, London. Aslib.

Nagel, Silke, Susanne Hezel, Katharina Hinderer, and Katrin Pieper, editors. 2009. *Audiovisuelle Übersetzung. Filmuntertitelung in Deutschland, Portugal und Tschechien*. Peter Lang Verlag, Frankfurt.

Och, Franz Josef and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Ortiz-Martínez, Daniel, Ismael García-Varea, and Francisco Casacuberta. 2005. Thot: A Toolkit to Train Phrase-Based Statistical Translation Models. In *Tenth Machine Translation Summit*, Phuket. AAMT.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing 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.

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

Popowich, Fred, Paul McFetridge, Davide Turcato, and Janine Toole. 2000. Machine Translation of Closed Captions. *Machine Translation*, 15:311–341.

Prokopidis, Prokopis, Vassia Karra, Aggeliki Papagianopoulou, and Stelios Piperidis. 2008. Condensing Sentences for Subtitle Generation. In *Proceedings of Linguistic Resources and Evaluation Conference (LREC)*, Marrakesh.

Specia, Lucia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the Confidence of Machine Translation Quality Estimates. In *Proceedings of MT-Summit*, Ottawa, Canada.

Tiedemann, Jörg. 2007. Improved Sentence Alignment for Movie Subtitles. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Volk, Martin and Søren Harder. 2007. Evaluating MT with Translations or Translators. What is the Difference? In *Proceedings of Machine Translation Summit XI*, Copenhagen.

Volk, Martin. 2008. The Automatic Translation of Film Subtitles. A Machine Translation Success Story? In Nivre, Joakim, Mats Dahllöf, and Beáta Megyesi, editors, *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, volume 7 of *Studia Linguistica Upsaliensia*, pages 202–214. Uppsala University, Humanistisk-samhällsvetenskapliga vetenskapsområdet, Faculty of Languages.