SUMAT: Data Collection and Parallel Corpus Compilation for Machine Translation of Subtitles

Volha Petukhova[1], Rodrigo Agerri[2], Mark Fishel[3], Yota Georgakopoulou[4], Sergio Penkale[5], Arantza del Pozo[1], Mirjam Sepesy Maučec[6], Martin Volk[3] and Andy Way[5]

[1]Human Speech and Language Technologies, Vicomtech-IK4, San Sebastiàn, Spain
[2]IXA NLP Group, University of the Basque Country, San Sebastiàn, Spain
[3]TextShuttle GmbH, Affoltern am Albis, Switzerland
[4]Deluxe Digital Studios, London, United Kingdom
[5]Applied Language Solutions Ltd., Manchester, United Kingdom
[6]Laboratory for Digital Signal Processing, University of Maribor, Slovenia

[1]{vpetukhova,adelpozo}@vicomtech.org; [2]rodrigo.agerri@ehu.es;
[3]{fishel,volk}@cl.uzh.ch; [4]yota.georgakopoulou@bydeluxe.com;
[5]{andy.way,sergio.penkale}@appliedlanguage.com; [6]mirjam.sepesy@uni-mb.si

Abstract

This paper describes the data collection and parallel corpus compilation activities carried out in the FP7 EU-funded SUMAT project. This project aims to develop an online subtitle translation service for nine European languages combined into 14 different language pairs. This data provides bilingual and monolingual training data for statistical machine translation engines which will semi-automate the subtitle translation processes of subtitling companies on a large scale.

Keywords: parallel multilingual corpora, statistical machine translation, subtitle translation service

1. Introduction

Subtitling plays an important role being the preferred method of translating multimedia content in most European countries and for most genres, making audiovisual content widely accessible across languages. Current European policy (European Commission, 2010) promotes the subtitling contents broadcasted by public TV networks. As a consequence, the demand for subtitling by the European audiovisual industry has increased in recent years (MCG, 2007).

However, subtitling faces some important problems that are preventing the expansion of the market such as cost, time and quality. Subtitling and audiovisual translation have been recognized as areas that could greatly benefit from the introduction of Statistical Machine Translation (SMT) followed by post-editing, in order to increase productivity (see e.g., Volk, 2008; Hardmeier and Volk, 2009; de Sousa et al., 2011).

Currently, there are no effective tools or services that can provide automatic subtitle MT services. The main limitation is the lack of sufficient parallel subtitle corpora required to train the SMT models.[1] Such data is available for some language pairs, e.g., in the OPUS OpenSubtitle corpus (Tiedemann, 2009), but being based on openly available subtitles with no quality checking, their usefulness has yet to be determined.[2]

Professionally produced high-quality subtitle data is the property of subtitling companies. Moreover, data is used and stored in various subtitle formats, some of which are proprietary, e.g. .o32, .x32 and .s32 belong to Softel, .890 is the Cavena format, .pac is the Screen format, and .ezt belongs to EZTitles. All this makes access to high-quality data for research and development purposes rather problematic.

The European project SUMAT[3] aims to develop an online subtitle translation service for nine European languages, combined into 14 different language pairs, in order to semi-automate the subtitle translation processes of subtitling companies on a large scale.

In order to obtain high-quality subtitle data, SUMAT has professional subtitle translation companies as members of the consortium, including Deluxe Digital Studios[4], Voice&Script International[5], InVision Ondertiteling[6] and Titelbild Subtitling and Translation.[7] Professional data of high quality, however, cannot be directly used as SMT training material: a parallel corpus should be compiled first. The problems that need to be solved are how to deal with different data formats and encodings, different formatting styles and subtitle cuts, differences in language structures, but also handling those human errors that occur when identifying and aligning parallel documents and subtitles. In this paper we describe what data has been collected (Sec-

[1]It is known from previous experiments reported in the literature (Hardmeier and Volk, 2009) that 7-10 million words (approximately 1 million subtitles) is a good size training corpus per language pair.

[2]The same applies to the new version of OpenSubtitles, which is to include much larger amounts of data and is currently under development.

[3]An Online **S**ervice for **SU**btitling by **MA**chine **T**ranslation: http://www.sumat-project.eu

[4]http://www.bydeluxe.com

[5]http://www.vsi.tv

[6]http://www.ondertiteling.nl

[7]http://www.titelbild.de

tion 2) and how the parallel corpus for training SMT systems has been compiled and evaluated (Section 3). Section 4 summarizes the results and outlines further SUMAT corpus developments.

2. Data collection

2.1. Corpora

The SUMAT subtitle translation service providers are major European and world players that have large subtitle resources available in more than 50 languages, produced by professionally trained subtitle translators who are native speakers of the target language employing multi-level quality control procedures. The SUMAT subtitling companies have specified the quantities and characteristics of the subtitle data and provided subtitle training material to be used in the project. Table 1 gives an overview of provided subtitles for each of the SUMAT language pairs.

Language pair	Amount of parallel subtitles
English - German	1.935.829
English - French	1.065.931
English - Spanish	848.318
English - Dutch	838.463
English - Swedish	635.804
English - Portuguese	560.716
Serbian - Slovenian	175.097
Total	6.060.158

Table 1: *Amount of available parallel subtitles provided by the members of the SUMAT consortium.*

Additional monolingual subtitle data for Dutch, English, German, French, Swedish and Portuguese has also been collected, in order to be used for language model training (see Table 2).

Language	Amount of monolingual subtitles
English	1.891.677
German	1.958.171
French	1.060.885
Dutch	2.609.869
Swedish	3.147.588
Portuguese	1.547.372
Total	12.215.562

Table 2: *Amount of available monolingual subtitles provided by the members of the SUMAT consortium.*

A sufficient amount of subtitles are available to train state-of-the-art SMT systems for all the SUMAT language pairs except for Slovenian–Serbian. In order to mitigate the relative scarcity of the training data for the Slovenian–Serbian language pair, special focus will be paid to (a) constraining the subtitle genre/domain, (b) compiling as many freely available parallel and monolingual non-subtitle data as possible, and (c) exploiting the linguistic similarity between the two languages, by making use of dictionaries, morpho-syntactic information, hierarchical and phrasal lexicons which have been used effectively for languages like Serbian in the past (see Popović and Ney, 2005; Popović and Ney, 2006).

Since SMT improvements can be expected through larger training sets, additional resources from freely available non-subtitle parallel corpora may be used to augment the subtitle data, e.g. the EuroparlTV parallel corpus, subtitles produced for the web television of the European Parliament,[8] the JRC-Acquis Multilingual Parallel Corpus[9] (a collection of legislative texts), or the ParaSol corpus of Slavic and other languages.[10] In order to efficiently use additional non-subtitle parallel data, the automatic extraction of sentence pairs from these corpora with similar characteristics (e.g. sentence length, limited grammatical complexity) with subtitles will be investigated. While these freely available non-subtitle corpora will help reduce the number of out-of-vocabulary items, their effect on overall MT quality is as yet unproven.

2.2. Data quality

The SUMAT subtitle providers are professional companies specialized in producing subtitle translations of the highest quality. They work exclusively with professionally trained translators and experienced freelancers (masters in translation, audiovisual translation and/or university degree in language studies) who are native speakers of the target language. Personnel receive subtitling training, and via a constant feedback cycle it is ensured that the subtitling and translation skills of all staff members are up-to-date and remain at a very high level.

In addition to the translation personnel, all SUMAT subtitling companies have a pool of experienced editors, reviewers and proofreaders. All work goes through four levels of quality control: (i) origination; (ii) review; (iii) proofreading; and (iv) technical quality control which includes verifying the number of subtitles, blank subtitles, blank rows, italics, positioning on the screen, double spaces, minimum intervals, timing violations, shot change timing, justification, treatment of on-screen text, etc. The technical quality control on the translation (i.e. text) aspect includes the following elements:

1. the language used is indeed the correct one.
2. the correct font is used to avoid character corruptions.
3. the correct font size is used (where applicable).
4. the subtitles do not exceed two lines (unless exceptions stated in the guidelines) and that the character limit per line is respected (where applicable).
5. all text has been translated and if not this is done for a good reason (e.g. as often happens in Dutch subtitles for English audio files).
6. all numbers have been correctly translated .
7. all proper names have been consistently treated.
8. dialogues in subtitles have been treated as per the company/client guidelines, e.g. use of dialogue dash at the start of each subtitle line and no more than two speakers per subtitle.
9. the appropriate use of punctuation and capitalisation.

[8] http://www.europarltv.europa.eu

[9] http://langtech.jrc.it/JRC-Acquis.html

[10] http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/

Genre	EN-NL	EN-FR	EN-DE	EN-PT	EN-ES	EN-SV	SR-SL
Series	378.633	369.779	266.565	229.068	338.749	43.082	6.185
Features	102.655	196.698	176.316	132.769	169.388	53.411	0
DVDextra	82.488	111.605	84.437	64.023	107.781	1.251	0
Other	35.509	109.699	130.406	676	62.069	378	2.821
Documentary	17.479	38.233	33.545	4.704	32.151	1.410	5.357
DVDCommentaries	12.671	29.324	28.178	50.361	0	0	0
News	18.898	2.618	48.431	0	0	0	37.930
SitComs	28.648	22.339	10.214	37.888	0	0	0
Corporate	3.828	41.172	33.623	602	0	0	0
Music	2.718	10.106	7.157	1.316	3.895	0	0

Table 3: *SUMAT parallel subtitle data subdivided into genres.*

Genre	EN-NL	EN-FR	EN-DE	EN-PT	EN-ES	EN-SV	SR-SL
Action	6.120	14.387	10.763	51.482	14.037	3.392	0
Adventure	21.654	10.113	12.682	2.186	10.088	0	0
Comedy	117.718	137.441	56.993	80.703	98.539	51.282	0
Culture	63.657	120.642	107.080	55.612	87.897	7.175	6.185
Daily	14.444	9.939	40.772	0	969	0	43.636
Drama	245.114	266.791	147.043	147.799	218.756	25.583	0
History	32.667	17.848	24.927	25.125	35.253	0	0
Horror	19.544	14.917	11.768	24.835	13.437	3.456	0
Mystery/Detectives	47.579	45.728	42.477	6.132	34.911	2.874	0
Nature	0	705	1.194	0	305	0	0
None	35.718	35.741	18.610	42.380	3.895	0	0
Other	47.045	192.464	231.531	38.063	164.540	378	58.758
Romance	10.608	16.512	7.242	7.308	12.501	3.694	0
ScienceFiction	16.717	16.375	16.228	2.669	15.510	1.698	0
Sports	818	603	53.078	33.283	34.240	0	1.849
Technology	3.505	36.940	35.950	2.573	34.660	0	0

Table 4: *SUMAT parallel subtitle data subdivided into domains.*

10. that there are no reading speed violations.
11. the line breaks are correct, i.e. segmentation is done at the highest possible syntactic node.

There are no official universal subtitling guidelines, but rather rules of thumb that are accepted in general by the subtitling industry but which may nonetheless be different from country to country. They differ, however, normally in details rather than the core principles of subtitling (see ITC Guidance on Standards for Subtitling, 1999[11]). Some general guidelines for subtitling can be found in documents of the European Broadcast Union (EBU), e.g. EBU Report Access Services that include such recommendations[12]. All companies though have their own internal subtitling guidelines and sometimes clients have too (Díaz-Cintas and Remael, 2007).

As for error ranking for subtitles in general, errors can be related to timing and to text. In case of timing issues, a list of potential errors in priority ranking are as follows:

1. The subtitles are out of synchronization with the audio.
2. Any type of timing violation which would either make the subtitles not display properly (e.g. overlapping subtitles) or would violate the company or client guidelines (e.g. minimum and maximum duration).
3. Minor errors would be subtitles not being entirely frame accurate, but viewers rarely notice such errors.
4. A type of error that can be more or less serious, depending on the extent of violation, is reading speed violation, which means that the text in the subtitles is too long for it to be read in the time allocated for them. This can either be considered a timing error (if the timings can be improved) or a textual error (if the timings are fine, but the text has not been condensed enough).

In case of text, a list of potential errors in priority ranking is as follows:

1. Missing translations, when it is important to the comprehension of the story that there be a translation provided (sometimes background dialogue does not have to be subtitled, or if it is deliberately left untranslated it does not constitute a grave error necessarily).
2. Mistranslations, i.e. anything that changes the meaning of the utterance.
3. Mistakes in grammar, syntax, spelling or punctuation can be grave or not depending on the result of the mistake, i.e. the extent to which it inhibits comprehension of the utterance.
4. Even if the subtitle is otherwise perfect, if there are reading speed violations (i.e. the text is less edited than it should be given the time allocated to each particular subtitle), then the subtitle is still problematic. The gravity of the error depends on the extent of the violation.

[11] Available via `www.ofcom.org.uk`

[12] For more information visit `http://tech.ebu.ch/publications`

5. Any violations of guidelines or inconsistencies that do not otherwise impede comprehension.

The quality of the automatically translated subtitles will be assessed in the translation experiments and the errors that come out of the machine pre-processing/translation of subtitles will be ranked. It is also planned to compare the amount of time it would take subtitling companies to originate a file from scratch as opposed to post-editing a pre-processed file so that it reaches the same quality level as a file originated from scratch.

As to the quality of the previously mentioned freely available parallel subtitle data that may be used in the project, the situation is different. As we noted in Section 1, the OpenSub corpus contains subtitle files that are produced by volunteers and which may include linguistic mistakes of various kinds (e.g. spelling, grammatical, semantic and stylistic). This data may be used only for those language pairs for which the need for additional training material is identified, and in any case only after careful selection and quality checking. Regarding the quality of the parallel non-subtitle data that may be used in SUMAT as an additional resource, the majority of the data available is of high quality. Translations and sentence alignments for Europarl, the JRC-Acquis Multilingual Parallel Corpus, EMEA, the ParaSol corpus and MULTEXT-East cesDoc have either been manually performed by experts or automatically with manual correction. The same is applicable to the EuroparlTV corpus, where subtitles and translations are of high quality, but created for parliamentary business, meetings and news, and when building translation and language models this should be taken into account.

2.3. Genres and domains

It has been noted in the literature (Volk, 2008) that the quality of automatically translated subtitles may vary considerably across film genres and that evaluation scores are rather domain-dependent. For this reason, in SUMAT subtitling companies provided information about genre and domain for each subtitle document. In total 16 genres have been defined (e.g. features, news, documentary), sub-divided into 21 domains (e.g. medicine, culture, history). Translation experiments will determine which strategy is better: either to have separate SMT systems per genre and domain or not. The amount of subtitle data available per genre and domain is computed. For those cases where it is impossible to provide domain information – because it might not be tractable or difficult to decide upon – documents are classified as 'other' and automatic domain classification techniques such as Support Vector Machines (Sharoff, 2007) will be applied. For those genres and language pairs that are under-represented, some genre classes will be merged and additional data from external open resources will be used. Tables 3 and 4 provide an overview of parallel subtitle data available per genre and domain, respectively.

3. Data pre-processing

In order to be used for SMT, data delivered by subtitling companies first needs to be pre-processed. This includes (1) subtitle format conversion, and (2) pre-processing of converted data.

The first task is concerned with the conversion of subtitle formats into plain text. Since some formatting information may be lost in the course of the conversion from some subtitle formats into plain text, a pivot format is needed to preserve formatting information throughout the translation process.

The second task is concerned with the pre-processing of the converted subtitle data. This involves the following steps: (a) alignment of parallel documents (for parallel corpora); (b) language identification (error checking); (c) tokenization; (d) normalization; (e) sentence splitting; (f) sentence alignment; and (g) subtitle alignment.

3.1. Data conversion and unicode normalization

As it has been already noted, subtitle formats are of various types, with more than 200 different subtitling formats in existence, some of them proprietary. For other formats, such as .stl and .xml, detailed specifications are publicly available.

In practice, subtitle files are stored in a mix of different formats and the online subtitle translation service to be developed has to be able to handle them. Therefore, we need to build subtitle format converters to plain text. The following subtitle formats will be supported by the SUMAT pilot translation service: EBU STL, EBU TT, TXT and SRT. These formats have been chosen because (a) they all are non-proprietary formats; (b) EBU STL is the current non-proprietary standard; (c) EBU TT will be the next improved standard; [13] (d) everyone can produce/convert to TXT format; and (e) SRT is becoming more widely employed because it is the format supported by youtube and webplayers in general. It will be possible to have each of these formats as input and/or output. This means that cross-conversion will be supported, e.g. the input file can be in TXT format and the output file in EBU TT format.

Text files delivered by subtitling companies are encoded using different character encoding standards. As a result, the format converters being developed include encoding detection. They can distinguish between UTF-8, UTF-16, Windows-1251 and Latin1, and employ UTF-8 as output encoding, since most of the tools to be used in SUMAT are compatible with such character encoding.

3.2. Dealing with formatting

Some subtitle file formats may contain formatting information. The most common formatting information includes positioning, coloring and italics. Positioning refers to the position of subtitles on the screen, e.g. if the file is for DVD purposes, subtitles are centered with subtitles containing dialogue centered and left-aligned. Coloring is used to encode speaker identity in teletext files, while italics emphasise special text such as off-screen narration.

[13] EBU TT stands for EBU Timed Text. It is an XML-based subtitling format intended as a follow-up to the EBU STL format. For more detailed information see EBU TT Part 1 – Subtitling format definition (EBU Tech 3350) – released in January 2012 for industry comments, as well as the XML schema of EBU TT Part 1, which is available for download at `http://tech.ebu.ch/ebu-tt`

While formatting should be saved and retained during translation within the online service, we will explore whether it needs to be eliminated during corpora compilation for SMT training. For example, (Du et al., 2010) demonstrated that maintaining the formatting as part of the SMT phrase-table actually improved MT quality. Files in .o32, .x32, .s32 and .stl do generally contain some positioning and coloring information; .srt files may have some italics and coloring information; and files in .txt have no formatting encoded. EBU TT has a detailed specification to encode all kinds of formatting information.

Among the formats to be supported by the SUMAT online pilot service, only EBU STL, SRT and EBU TT will contain formatting information consistently. When the input file is in any of these formats, the system will have to keep track of its formatting. A pivot format is needed to preserve formatting information throughout the translation process within the online service. Such a pivot format will be EBU TT, as it is capable of preserving all formatting information in an unambiguous way, all its formatting information is encoded in terms of easily parseable XML attributes.

3.3. Language identification and document alignment

For SMT training purposes, the language of the input subtitle files needs to be checked and parallel files need to be aligned and assigned the same name ID. Subtitling companies provided this information when delivering their data. However, language and name IDs sometimes contain errors. Thus, language identification is required to double-check the original language classification provided by the subtitling companies for erroneous language assignment. For language identification, SUMAT uses the *Lingua::Ident* tool.[14] Lingua::Ident is a statistical language identifier based on *n*-gram probabilities for languages and is open source. Only 0.12% of the files were identified as being not one of the SUMAT languages and were replaced with the correct versions.

Documents that contain subtitles in different languages (source and target language) that are translations of one another are parallel. One of the ways to detect parallel subtitle files is to extract the time codes that specify a subtitle's start and end time for each document and measure the correspondences in time codes with all other documents.

It was originally assumed that subtitling companies could guarantee that at least 90% of the time-codes of corresponding parallel subtitle files would be identical within a threshold of seven film frames. However, a number of difficulties were experienced when performing document alignment based on time code correspondence, e.g. differences in the time codes for some parallel subtitle files greater than seven frames were detected. Moreover, even if less than 90% of the time codes match, documents may still be parallel. This often happens when one language version may have an offset due to a longer introduction or a different cut, or when some subtitles are not translated and the timeline shifts. Time codes do not correspond in files that were not created as translations of one another, but from a different source file.

[14] `http://search.cpan.org/\~mpiotr/Lingua-Ident-1.6/Ident.pm`

Language pair	Amount of parallel subtitles (converted)
English - German	1.358.010
English - French	987.935
English - Spanish	811.171
English - Dutch	801.529
English - Swedish	594.505
English - Portuguese	545.217
Serbian - Slovenian	169.654
Total	5.268.021

Table 5: *Amount of available parallel converted subtitles.*

This was taken into consideration when designing and implementing the algorithm for document alignment, which also matches shifted documents automatically based on dynamic programming.

Table 5 shows the results of parallel data format conversion, language identification and document alignment. When compared with the delivered data (see Table 1), about 12% of the parallel data in terms of subtitles were lost during the conversion, language identification and document alignment stages. As far as non-parallel documents are concerned, these were added to the monolingual data collections, with 13,663,880 monolingual subtitles in total being converted: for Swedish, 3,192,674; English 2,712.442; Dutch 2,644,780; German 2,335,948; Portuguese 1,556,942; Spanish 49,540; Serbian 19,503; and Slovenian 5,850.

3.4. Tokenization and normalization

Subsequently, subtitle data needs to be normalized (i.e lowercased) and tokenized (i.e. split into a set of tokens which constitute atomic parseable elements such as words, abbreviations, acronyms or punctuation marks). For both tasks, the available Moses scripts (Koehn et al., 2007)[15] were used and evaluated, and the necessary extensions to adapt them for subtitle processing were developed. As for lowercasing, this is one of the simplest pre-processing tasks, where capital letters were substituted by their corresponding lower cases. Since capital letters were used as features for later pre-processing tasks (e.g. sentence splitting), this task was performed as the last one in the SUMAT pre-processing pipeline. The performance on this task was 100% correct.

As for tokenization, when tokenizing text for MT, the main guiding principle is to reduce text to a sequence of tokens from a small inventory. It is not required, for example, to learn different translations for *house*, depending on whether it is followed by a comma *'house,'* or surrounded by quotes *"house"* or brackets *(house)*. Similarly, the sparseness of data should be reduced in the way that, for example, "$" is always translated into Spanish as *dólares* independent of whether it was "$60" or "$5". In the same way, measurements like *km/h* or $^{\circ}C$ are better to be separated from the numbers. Accordingly, splitting off punctuation, possessive or plural markers, apostrophes in merged words (e.g. *don't*), hyphens and other special symbols is the main tokenization issues for the most European languages. Punctuation marks may, however, be part of a token (for ex-

[15] `http://www.statmt.org/moses/`

ample when dealing with abbreviations). Since subtitling companies do not use special lists with abbreviations for SUMAT languages, lists of abbreviations were compiled for each language from available general lists. A list of Slovenian abbreviations was extracted from a national corpus FidaPLUS;[16] for English, the Oxford dictionary list and those of the Encyclopedia Britannica were used; for German, the available list was supplemented using the web databases of the common German abbreviations,[17] and similar approaches were used for French, Spanish and Dutch.[18] General abbreviation lists were supplemented with special cases encountered in subtitle domain. The SUMAT data has been tokenized and contains 86,952,053 tokens in the parallel corpora, and 136,625,878 tokens in the monolingual ones.

The performance on the tokenization task was evaluated by assessing accuracy, precision and recall. Accuracy was calculated using the following formula: $(tp + tn)/(tp + tn + fn + fp)$, precision $= tp/(tp + fp)$ and recall $= tp/(tp + fn)$, where *tp* is the number of correctly tokenized tokens; *tn* is the number of strings of characters that are not tokens but are correctly recognized as such; *fp* is the number of incomplete tokens, e.g. *'92* is one token and is recognized as two tokens such as *' 92*, where *Mr.* is tokenized as *Mr .* or where *o'clock* becomes *o 'clock*; and *tn* is the number of strings of characters that form more than one token but are wrongly recognized as just a single token, e.g. 5° C is tokenized as $5\,^\circ$C. For this, 500 subtitles per language pair were randomly selected as testing material provided that this set contained data of the different subtitling companies balanced per genre and domain. Subsequently, the selected material was manually tokenized, sentence split and aligned by experts (linguists) so as to construct the gold (or reference) data sets against which to compare the test data. The overall performance on the tokenization task is good, with subtitles tokenized with high accuracy and precision (0.99 on average) and minimum token loss (recall 0.98 on average).

3.5. Parallel data alignment

Compilation of any parallel corpus requires alignment of various types. For SUMAT purposes at the pre-processing stage, the following alignments were performed: (1) sentence alignment to align translations at the sentence level; and (2) subtitle alignment to identify parallel subtitles.

3.5.1. Sentence splitting and alignment

Previous work on SMT of subtitles (e.g. Armstrong et al., 2006; Volk, 2008) has developed successful systems without splitting subtitles into sentences, with this approach working particularly well for closely related languages. Accordingly, we decided to use this strategy within SUMAT. However, we aim to use additional translation models using the linguistically annotated subtitle data to create factored models in Moses. Linguistic annotation usually operates on the sentence level. Thus, subtitles that may contain one or more sentences will need to be split into separate sentences. There are many freely available tools for this task, e.g. OpenNLP tools.[19] However, the majority of the sentence splitting tools are for English. Therefore, in SUMAT a tool was developed adopting a language-independent approach based on sentence-final punctuation marks as identified during tokenization. Recognizing the end of a sentence is not always a trivial task. Punctuation marks that usually appear at the end of a sentence may not indicate the end of a sentence, e.g. the dot can be the abbreviation marker as well. This problem cannot be solved by using abbreviation lists, since the dot is then ambiguous (however, very few cases were observed where the dot is part of the abbreviation and sentence-final punctuation mark at the same time). Additionally, the use of multiple dots '...' is ambiguous: they often occur in the middle of a sentence indicating pauses in speech and at the end of the sentence when the sentence is not finished (e.g. incomplete abandoned dialogue utterances). The latter problems were largely solved by taking capitalization and other punctuation marks into account, e.g. *'Y dejemos claro ... lo que no vamos a hacer'* was merged together in one sentence, and *'- Pero ...'* *'- Pero qué ?'* split into two sentences. Missing periods at the end of a sentence caused some problems. For instance, lyrics (songs) have no sentence-final punctuation and every line starts with capital letter, with 18 cases having been found so far in the EN–ES data. For such cases, problems were solved by keeping line breaks for files from the Music domain/genre. In some cases, however, sentence-final punctuation marks (mostly a full stop) are missing because of human error, but such cases are rare. Table 6 presents the number of identified and aligned sentences in the SUMAT corpus.

The performance on the sentence-splitting task for SUMAT data was evaluated. The selection procedure was slightly different from those for tokenization evaluation: rather than using a randomly selection of subtitles, consecutive subtitles were selected to construct the test and reference sets. Furthermore, accuracy, precision and recall was computed as described above, but the interpretation is different from the tokenization task: *tp* is the number of correctly split sentences; *tn* is the number of tokens that while not forming a sentence are correctly recognized as such, e.g. incomplete abandoned sentences; *fp* is the number of incomplete sentences wrongly identified as a full sentence (< 1 sentence); and *tn* is the number of more than one sentence wrongly spanned in one sentence (> 1 sentence). The overall performance on this task is good, as we obtained an accuracy of 0.97, precision 0.96, and near perfect recall of 0.99.

Sentences from one language and their translation were identified and mapped. For sentence alignment two approaches were used: text-independent based on time code information, and a text-based approach. For the former we

[16] www.fidaplus.net

[17] See, for example, http://www.abkuerzungen.org

[18] http://spanish.about.com/od/writtenspanish/a/abbreviations.htm, http://www.wordreference.com/es/Abbreviations-Spanish.aspx and http://en.wiktionary.org/wiki/Category:Dutch_abbreviations

[19] http://incubator.apache.org/opennlp/index.html

Language pair	Number of sentences identified		Aligned sentences
	source language	target language	
English-Dutch	972.135	922.602	702.877
English-French	1.022.379	1.000.177	883.371
English-German	1.084.489	1.186.049	768.471
English-Portuguese	603.369	598.554	545.995
English-Spanish	810.814	798.259	722.317
English-Swedish	732.323	705.140	635.723
Serbian-Slovenian	211.068	209.426	110.481
Total	5.436.577	5.420.207	4.369.235

Table 6: *Number of identified and aligned sentences in the SUMAT corpus.*

developed a separate tool. For the latter approach, we used the Hunalign tool (Varga et al., 2005). Hunalign supports a mode in which bilingual dictionaries are automatically generated through sentence-length alignment. Then it realigns the text in a second pass, using the automatic dictionary. Both approaches were evaluated on the SUMAT data and the results are very encouraging. Both approaches result in good performance (average accuracy of 0.96, precision and recall of 0.97). Results were worse for the Serbian–Slovenian language pair, where the accuracy was 0.45, the precision was 0.81 and the recall 0.43, on alignment using time codes. In order to improve performance, data was chosen that was aligned using Hunalign, where the performance was significantly better: accuracy was 0.74, precision 0.74 and recall was very high at 0.99. The errors identified are concerned with the fact that most of the Serbian–Slovenian data are not direct translations of one another, but from different source files, therefore problems appeared such as (i) the different lengths of source and target sentences, where in general, short sentences were better aligned; and (ii) 1-to-n sentence correspondence, which HunAlign cannot handle, while alignment based on time codes was quite successful in dealing with these problems. Table 6 gives an overview of identified and aligned sentences for the SMT system training.

3.5.2. Subtitle alignment

Parallel subtitles are identified and aligned using time codes, i.e. assuming that subtitle files with (almost) identical time codes contain parallel subtitles. However, some inconsistencies were detected when performing this type of alignment as described in Section 3.3. Due to different cuts in source- and target-language files, time code-based alignment is far from perfect. In alignment based on time-codes, some files with 1-to-1 subtitles had skewed time-codes on one side, e.g. the first subtitle time-codes are equal, the 10th subtitle time-codes are 2 seconds apart, the 100th subtitle time-codes are 20 seconds apart, etc. This also demanded an algorithm of its own to handle such problems. For small time stamp shifts, different tolerance values were used for different language pairs, e.g. up to 1 sec (25 frames) for English–Spanish and up to 1.05 sec for English–Swedish and Slovenian–Serbian.

We also performed text-based subtitle alignment and compared both algorithms. Additionally, a subtitle may contain dialogues between two different speakers. We also investigated whether merging speaker lines or not helps improve the alignment of parallel subtitles for some languages. This was checked on three language pairs, and the performance on alignment was indeed slightly better for English–Spanish and English–French, and for Serbian–Slovenian when using Hunalign.

PARALLEL CORPORA	Number of aligned subtitles
English-Dutch	687.879
English-French	947.141
English-German	829.858
English-Portuguese	523.417
English-Spanish	784.825
English-Swedish	589.338
Serbian-Slovenian	112.293
Total	4.474.752

Table 7: *Number of aligned subtitles in the SUMAT corpus.*

The performance of both tools on different types of subtitle cut was evaluated. Alignment using time codes was more accurate with an accuracy of 0.89 on average (compared to 0.8 using Hunalign), precision of 0.94 (Hunalign: 0.82) , and recall of 0.91, slightly lower than the 0.96 obtained with Hunalign. The results obtained are reasonably good and can be considered satisfactory, since the results reported in the literature (e.g. Tiedemann, 2009) are significantly lower. For similar reasons as above, the results are again worse for Serbian-Slovenian data, where the best performance using time codes is: accuracy is 0.47, precision - 0.74, and recall - 0.49; using Hunalign the accuracy reached is 0.51, precision - 0.54, and recall - 0.78. Here, time code-aligned data will be chosen for training, since the alignment is not only relatively accurate, but also more precise which is important for MT.

Table 7 gives an overview of aligned parallel subtitles in the SUMAT corpora. As expected, some parallel data (about 15%) was lost in the course of alignment (compared with the delivered data presented in Table 1).

4. Conclusions and future work

In this paper we discussed the data collection and parallel corpus compilation process for training SMT systems, which includes several procedures such as data partition, conversion, formatting, normalization and alignment. The paper illustrates how even with the availability of professionally produced data, sentence and subtitle alignment is far from a trivial task. There are many issues that need to be taken into account. We discussed in detail each data pre-processing step using various approaches. We also presented evaluation results and statistics on the final SUMAT

subtitle corpus, as well as a discussion on the data preprocessing procedures. We experienced a loss of approximately 12% of the parallel data in terms of subtitles during the conversion, language identification and document alignment stages. Some parallel data was lost in the course of subtitle alignment (about 15%). The total loss of parallel data was about 27%, but the situation per language pair is different, e.g. for English–German the data loss due to document alignment problems (1-to-n correspondence) was the highest (about 57%), followed by Serbian–Slovenian due to skewed time codes in target and source files (loss of 35%). For English–Spanish, English–Portuguese and English–Swedish, the data loss was the lowest (about 7%), while for English–Dutch and English–French the loss was about 14%. Not-aligned data was added to the monolingual data.

The collected corpus size for subtitles to be used as SMT training material for all language pairs is sufficient, and it is close and for some language pairs exceeds the ideal size (around 1 million subtitles) required for such a task (see (Hardmeier and Volk, 2009). Apart from the quantity, the SUMAT corpus has a number of very important characteristics. First of all, high quality both in terms of translation and in terms of high-precision alignment of parallel documents and their contents has been achieved. Secondly, the contents are provided in one consistent format and encoding. Finally, additional information such as type of content in terms of genres and domain is available.

Language pair	Extra parallel subtitles	Total finals
English - German	346.500	1.704.510
English - French	500.500	1.488.435
English - Spanish	139.500	950.671
English - Dutch	614.500	1.416.029
English - Swedish	323.500	918.005
English - Portuguese	202.000	747.217
Serbian - Slovenian	40.000	209.654
Total	2.166.500	7.434.521

Table 8: *Amount of extra parallel subtitles that will provided by the SUMAT subtitling companies.*

The SUMAT parallel subtitle corpus has required substantial investment. We expect it to have a great impact on the rest of the project. The SUMAT project consortium will continue to maintain the corpus and to take an interest in its growth; extra data will be delivered by SUMAT subtitle providers (see Table 8), which will be supplemented with data from other resources, e.g. EuroparlTV and the OpenSub corpus. In the future, after the SUMAT lifecycle, we will consider the possibility of making the SUMAT corpus generally available to the wider community for research purposes.

Acknowledgement

The work leading to the results reported in this paper has received funding from the European Community grant agreement N 270919. The authors are also very thankful to the anonymous reviewers for their valuable and constructive comments.

5. References

Armstrong, S., C. Caffrey, M. Flanagan, D. Kenny, M. O'Hagan and A. Way. 2006. *Leading by Example: Automatic Translation of Subtitles via EBMT*. Perspectives 14(3):163–184.

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

Du, J., J. Roturier and A. Way. 2010. *TMX Markup: A Challenge When Adapting SMT to the Localisation Environment*. In Proceedings of the 14th Annual Meeting of the European Association for Machine Translation, St. Raphael, France [no page numbers].

European Commission. 2010. *Audiovisual Media Services Directive (AVMSD - 2010/13/EU). Official Journal of the European Union, 10 March 2010*.

Hardmeier, C. and Volk, M. 2009. *Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles*. In: Proceedings of 17th Nordic Conference on Computational Linguistics (Nodalida), Odense, Denmark, pp. 57–64.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. *Moses: open source toolkit for statistical machine translation*. In: ACL 2007: proceedings of demo and poster sessions, Prague, Czech Republic, pp. 177–180.

Media Consulting Group. 2007. *Study on dubbing and subtitling needs and practices in the European audiovisual industry.* On behalf of the Information Society and Media Directorate General and the Culture Directorate General of the European Commission, November 2007.

Popović, M. and Ney, H. 2005. *Exploiting phrasal lexical and additional morpho-syntactic language resources for statistical machine translation with scarce training data.* In: 10th EAMT Conference: Practical Applications of Machine Translation, Conference Proceedings, Budapest, Hungary, pp. 212–218.

Popović, M. and Ney, H. 2006. *Statistical Machine Translation with a Small Amount of Bilingual Training Data*. In: Proceedings of the 5th LREC SALTMIL Workshop on Minority Languages, Genoa, Italy, pp. 25–29.

Sharoff, S. 2007. *Classifying Web corpora into domain and genre using automatic feature identification.* In: Cahiers du Cental **4**: 83–94.

de Sousa, S., Aziz, W. and Specia, L. 2011 *Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles*. In: Proceedings of Recent Advances in Natural Language Processing Conference (RANLP-2011), Hissar, Bulgaria.

Tiedemann, J. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces.* In: N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp. 237–248, John Benjamins, Amsterdam/Philadelphia

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. and Nagy, V. 2005. *Parallel corpora for medium density languages.* In: International Conference: Recent Advances in Natural Language Processing, Proceedings (RANLP 2005), Borovets, Bulgaria, pp. 590–596.

Volk, M. 2008. *The Automatic Translation of Film Subtitles. A Machine Translation Success Story?* In: Joakim Nivre, Mats Dahllöf and Beáta Megyesi (eds.): Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein. Volume 7 of Studia Linguistica Upsaliensia, Uppsala, Sweden, pp. 202–214.