# Recycling Texts: Human Evaluation of Example-Based Machine Translation Subtitles for DVD

Marian Flanagan, BA (Mod.), MA

A dissertation submitted to Dublin City University in fulfilment of the requirements for the award of

Doctor of Philosophy

School of Applied Language and Intercultural Studies
Dublin City University

Supervisor: Dr. Dorothy Kenny

June 2009

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:_____ ID No.:_____
(Candidate)

Date: _____

## Contents

## List of Figures

**Chapter 2**



**Chapter 3**



**Chapter 4**



**Chapter 5**

## List of Tables

**Chapter 1**



**Chapter 2**



**Chapter 3**



**Chapter 4**

**Chapter 5**

## List of Examples

**Chapter 2**



**Chapter 3**

## List of Abbreviations

| | |
|---|---|
| **ACL** | Association for Computational Linguistics |
| **AVT** | Audiovisual Translation |
| **CBMT** | Corpus-Based Machine Translation |
| **CELEX** | Centre for Lexical Information |
| **CESTA** | Campagne d'Evaluation de Systèmes de Traduction Automatique (Machine Translation Evaluation Campaign) |
| **DARPA** | Defence Advanced Research Projects Agency |
| **DVD** | Digital Versatile Disk |
| **EAGLES** | Expert Advisory Group on Language Engineering Standards |
| **EAMT** | European Association for Machine Translation |
| **EBMT** | Example-Based Machine Translation |
| **EWG** | Evaluation Working Group |
| **FEMTI** | Framework for the Evaluation of Machine Translation in ISLE |
| **HLT** | Human Language Technology |
| **HP** | Harry Potter |
| **HPC** | Harry Potter Corpus |
| **HOH** | Hard-of-Hearing |
| **HPC** | Harry Potter Corpus |
| **IATIS** | International Association for Translation and Intercultural Studies |
| **ICON** | International Conference on Natural Language Processing |
| **ISLE** | International Standards for Language Engineering |
| **ISO** | International Organization for Standardization |
| **IWSLT** | International Workshop on Spoken Language Translation |
| **LDC** | Linguistic Data Consortium |
| **LOTR** | Lord of the Rings |
| **LOTRC** | Lord of the Rings Corpus |
| **LREC** | Language Resources and Evaluation |
| **MGC** | Mixed general Corpus |
| **MT** | Machine Translation |
| **OCR** | Optical Character Recognition |
| **RBMT** | Rule-Based Machine Translation |
| **SDH** | Subtitling for Deaf and Hard-of-Hearing |
| **SL** | Source language |
| **SMT** | Statistical Machine Translation |
| **ST** | Source Text |
| **TL** | Target language |
| **TM** | Translation Memory |
| **TT** | Target Text |

## Acknowledgements

I would like to sincerely thank everyone who assisted in any way with this research.

First and foremost, I wish to thank my supervisor Dr. Dorothy Kenny for her unwavering enthusiasm, her excellent academic guidance, her constant encouragement and support, for listening to my ideas and for being extremely generous with her time during the PhD process. I would also like to thank her for providing much-needed travel assistance and for the opportunity to teach while completing the PhD.

I wish to thank Dr. Minako O'Hagan for giving me the opportunity to begin my PhD studies through the Enterprise Ireland Proof of Concept scheme (MovRat project). Throughout my time at DCU she has offered guidance and encouragement, has listened to my teaching ideas and was also very generous with her time. I am very grateful for the financial support from Enterprise Ireland during the first year of my studies.

I would also like to thank Prof. Andy Way and the members of the National Centre for Language Technology (NCLT) for providing the EBMT system used in this study and for their help with the many machine translation questions posed during the research.

I could not have continued my research without the financial support from the Irish Research Council for the Humanities and Social Sciences (IRCHSS) to whom I am particularly indebted. In addition I would like to thank the School of Applied Language and Intercultural Studies (SALIS) and the Office of the Vice-President for Research (OVPR) for providing invaluable travel funding.

I am very grateful to the SALIS staff members who participated in the evaluation, Dr. Annette Simon, Ms. Anna Weiss and Dr. Heinz Lechleiter. I also wish to thank the forty-four subjects who participated in the research, without whom, this research would not have been possible.

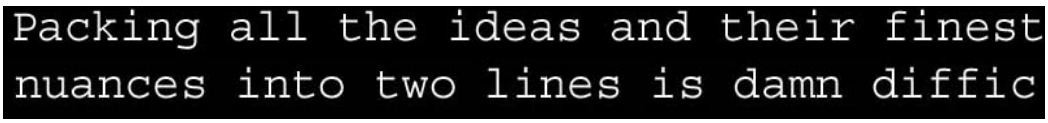
I am truly indebted to Mr. Gerry Conyngham, DCU Business School, for his help and advice with the many statistics-related questions, and for guiding me in the right direction.

I wish to thank my friends and extended family who have always been there to listen to me, to offer sound advice and to always know what to say to make me smile. In particular, a big thank you goes to my uncle, Dr. Patrick Flanagan, for proof-reading my thesis.

But most of all, I would like to thank my parents, Maireád and John, my sister, Lisa, my brother-in-law, Paul and my niece, Anna. Their constant encouragement, love, patience and support have made me realise my potential and I am extremely grateful.

And finally to Esben – Tak for at være en rigtig god ven, for opmuntringen, og mest af alt for at være dig. Du er virkelig noget særligt!

I dedicate this thesis to my grandparents and my parents for their constant encouragement and belief in me




Ivarsson (1992: front cover)

# Abstract


**Recycling Texts: Human evaluation of Example-Based Machine Translation subtitles for DVD**

**Marian Flanagan**

This project focuses on translation reusability in audiovisual contexts. Specifically, the project seeks to establish (1) whether target language subtitles produced by an EBMT system are considered intelligible and acceptable by viewers of movies on DVD, and (2) whether a relationship exists between the 'profiles' of corpora used to train an EBMT system, on the one hand, and viewers' judgements of the intelligibility and acceptability of the subtitles produced by the system, on the other. The impact of other factors, namely: whether movie-viewing subjects have knowledge of the soundtrack language; subjects' linguistic background; and subjects' prior knowledge of the (*Harry Potter*) movie clips viewed; is also investigated.

Corpus profiling is based on measurements (partly using corpus-analysis tools) of three characteristics of the corpora used to train the EBMT system: the number of source language repetitions they contain; the size of the corpus; and the homogeneity of the corpus (independent variables). As a quality control measure in this prospective profiling phase, we also elicit human judgements (through a combined questionnaire and interview) on the quality of the corpus data and on the reusability in new contexts of the TL subtitles. The intelligibility and acceptability of EBMT-produced subtitles (dependent variables) are, in turn, established through end-user evaluation sessions. In these sessions 44 native German-speaking subjects view short movie clips containing EBMT-generated German subtitles, and following each clip answer questions (again, through a combined questionnaire and interview) relating to the quality characteristics mentioned above.

The findings of the study suggest that an increase in corpus size along with a concomitant increase in the number of source language repetitions and a decrease in corpus homogeneity, improves the readability of the EBMT-generated subtitles. It does not, however, have a significant effect on the comprehensibility, style or well-formedness of the EBMT-generated subtitles. Increasing corpus size and SL repetitions also results in a higher number of alternative TL translations in the corpus that are deemed acceptable by evaluators in the corpus profiling phase. The research also finds that subjects are more critical of subtitles when they do not understand the soundtrack language, while subjects' linguistic background does not have a significant effect on their judgements of the quality of EBMT-generated subtitles. Prior knowledge of the *Harry Potter* genre, on the other hand, appears to have an effect on how viewing subjects rate the severity of observed errors in the subtitles, and on how they rate the style of subtitles, although this effect is training corpus-dependent. The introduction of repeated subtitles did not reduce the intelligibility or acceptability of the subtitles. Overall, the findings indicate that the subtitles deemed the most acceptable when evaluated in a non-AVT environment (albeit one in which rich contextual information was available) were the same as the subtitles deemed the most acceptable in an AVT environment, although richer data were gathered from the AVT environment.

# Introduction

The digital challenge is one of four new challenges Gambier (2008:25) believes the domain of Audiovisual Translation (AVT) is now facing, and this challenge is also central to the current research. The move from analogue technology to a much faster, reliable, flexible and compact digital technology in the 1990s has had, and will continue to have, a marked effect on the art of subtitling (ibid). Traditionally subtitling is a human-based process, but the role technology is playing within the domain is increasing (Díaz Cintas 2005). In the latter half of the 1970s dedicated subtitling equipment appeared on the market, and by the mid-1980s time codes were used to insert subtitles, revolutionising the process (Ivarsson 1992:25-26). However, even though computer-based subtitling systems are in use today, the production of subtitles from the spoken dialogue is practically unaided by any tools. The role of today's systems is to facilitate purely mechanical functions, including cueing the subtitles, spell-checking and other basic text processing functions (O'Hagan 2003a). Many of these subtitling practices were born during the age of analogue technology. However, digitalisation has meant significant changes for the subtitling industry, with one obvious example being the introduction of the DVD (Digital Versatile Disk). As O'Hagan (2007:157-158) points out:

> DVDs introduced interactivity and a degree of personalisation with the opportunity for the viewer to select the preferred mode of language support… [they have] contributed to the increased demand for subtitles in a wide range of languages, leading to a new approach to producing them within a limited time-frame and budget.

These new demands for subtitles on digital media have led O'Hagan (2003b) to investigate whether computer-aided translation (CAT) tools could be used to support (human) subtitlers. Others (e.g. Popowich et al. 2000, Piperidis et al. 2004, 2005, Melero et al. 2006, Armstrong et al. 2006c, Armstrong 2007, Volk & Harder 2007, Volk 2008, Hardmeier & Volk 2009), have turned their attention to Machine Translation (MT), in an attempt to investigate whether the translation of subtitles could be automated. As with most contemporary MT research, these sources have also been concerned with the evaluation of automatically produced subtitles. The current thesis is a contribution to this growing literature in the evaluation of machine translated subtitles.

More specifically, the thesis investigates whether subtitles generated by an Example-Based MT (EBMT) system are intelligible to and accepted by a DVD-viewing audience. It further seeks to establish whether the 'profile' of the corpus data on which the EBMT system draws affects intelligibility and acceptability of automatically generated subtitles. These ideas are expanded upon below. First, we give a brief indication of why subtitles might be amenable to automatic translation, we discuss developments that pave the way for increased automation in subtitle translation, and suggest ways of evaluating the success of automatic translation of subtitles.

## Characteristics of Subtitles

This growing interest in automating the subtitling process invites us to examine the characteristics of subtitles that are deemed to make their translation amenable to automation. Hardmeier & Volk (2009:1) mention that according to Becquemont (1996), "the characteristics of subtitles are governed by the interplay of two conflicting principles: *unobtrusiveness* and *readability*." This means that subtitles should allow viewers to understand the meaning of the dialogue without detracting from enjoyment of the movie. There are spatial and temporal constraints applied to subtitling, meaning subtitles tend to be short and written in a more simple form than sentences in printed sources. Díaz Cintas & Remael (2007:145) point out that such simplification is normal when one converts from the oral to the written, and note that "since the verbal subtitle sign interacts with the visual and oral signs and codes of the film, a complete translation is not required." They add that text reduction (partial or total) is commonly applied in subtitling, and the process of text reduction usually takes the form of eliminating information irrelevant for understanding purposes, followed by reformulating the relevant information in as concise a form as is possible or required (ibid:146). The rules governing subtitles are not set in stone and norms and conventions have evolved quickly (ibid:96). Ivarsson & Carroll (1998:67) comment that the norm for subtitles is that they span one or two lines, and contain up to a maximum of 80 characters (which results in an average reading speed of 175 words per minute). These reading speeds were developed for television audiences. With the advent of the DVD, norms and conventions have evolved once again and reading speeds are being increased. For example 180 words per minute is taken to be the norm with some companies applying even higher rates (Díaz Cintas & Remael 2007:98-99). Some subtitles can, of course be shorter than 80 characters. Hardmeier & Volk (ibid:2) for example, maintain that many

subtitles contain only two or three words to signify affirmation, negation or abuse. Brief subtitles can also be expected to be syntactically simple, exhibiting few if any instances of long-range dependencies (ibid). Such characteristics make automatic translation of subtitles easier. And even if some subtitles are ill-formed according to standard grammars, this is not particularly problematic given contemporary Corpus-Based Machine Translation (CBMT, see section 1.4.1) techniques, which do not in any case conduct linguistic analysis of input.

### Genesis Files

A recently emerging approach in order for human subtitlers to efficiently produce subtitles simultaneously in multilingual versions (and within short time frames) is known as the template-based approach (using so called "genesis files"). Subtitlers are provided with templates which include intralingual subtitles in the source language and time-codes. The subtitler is then required to fill in the target language subtitle translation. This process is usually conducted while the subtitler watches the movie being subtitled (Carroll 2004). Carroll (ibid) points out that such a template could make sense if it was thoroughly researched and well-timed. She adds that if subtitlers are free to use the template as an aid and are not compelled to force their translation into the template provided, this approach offers clear advantages. However, she also points out the possible disadvantages of this approach adding that the rigidity of such files can result in poor subtitling with little adherence to now common standards of good subtitling practice (cf. Ivarsson & Carroll 1998). Another issue is that sometimes subtitlers are required to fill in subtitle translations in these templates without actually watching the movie. The approach can be seen as the beginning of the profession moving towards (semi)-automating translation as it has a considerable standardising impact on subtitles across different languages because it provides the source input in a fixed manner (Minako O'Hagan: personal communication). This incipient technologisation motivates in part the current research and the use of machine translation in the production of subtitles.

### Machine Translation Evaluation Strategies

As already indicated, machine translation evaluation is at the heart of this thesis. It is now common practice to evaluate MT output using automatic metrics, as they are a quick, easy and cheap way to gauge the 'quality' of MT output (Papineni et al. 2002).

These metrics, however, are used normally in text-based domains which do not use other semiotic channels of communication such as sound and image. Many automatic metrics have been designed with the aim of gauging document-level reliability, and these scores have been shown to correlate well with human judgements. However, they have been criticised for inadequate correlation at sentence level (Callison-Burch et al. 2006). Given that subtitles differ in length and lexical variety to general text sentences (O'Hagan 2003b, Armstrong et al. 2006c), it is necessary to test the applicability of these metrics within the subtitling domain. In previous related studies automatic metric scores have been generated by Armstrong et al. (2006c) and more extensively by Armstrong (2007), Volk & Harder (2007) and Volk (2008). But even when automatic metrics are used in the MT research community, they still come second to human judgements of MT output, with automatic scores described as an imperfect substitute for human assessment of translation quality (Callison-Burch et al. 2007:139).

Three points emerge from the discussion so far. Firstly, it is not self-evident that purely text-based automatic metrics can be used to gauge quality in multi-modal communication. Secondly, some of the automatic metrics currently in use may not be optimally suited to environments where text is viewed (and possibly evaluated) on a segment-by-segment basis. Thirdly, whether or not automatic metrics can be said to correlate with human judgements of quality in audiovisual contexts is still something of an open question. To answer this question we need to conduct both human evaluations and automatic evaluation of MT output. To date, however there has been no comprehensive human-user evaluation of automatically generated subtitles. Of the studies to date that investigate the use of automated subtitling solutions, only Armstrong et al. (2006c) use end-users to evaluate the machine-generated subtitles. Popowich et al.'s (2000) study was conducted before the development of automatic metrics, and so evaluation is in this case necessarily human evaluation, but they used only one translator to evaluate the output. Armstrong et al.'s (ibid) pilot study tested different ways of implementing human evaluation of machine-generated subtitles, but this was done only on a small scale. The current study is thus, to our knowledge, the first to conduct a substantial human evaluation of machine-generated subtitles. We evaluate (in a real-use scenario) German subtitles produced by the Example-Based MT system MaTrEx, on the basis of three different subtitle corpora.

### Research Questions

This thesis aims to investigate two broad questions:

**RQ1**: Are EBMT-generated subtitles deemed intelligible and acceptable from the point of view of end-users of subtitled movies on DVD?

**RQ2:** Is there a relationship between the 'profiles' of corpora used to train an EBMT system, on the one hand, and viewers' judgements on intelligibility and acceptability of the subtitles produced by the system, on the other?

The corpus profiles referred to in RQ2 can be further broken down into three components:

- The number of SL repetitions in the corpus
- The size of the corpus
- The homogeneity of the corpus

The study is divided into two phases: during the first 'prospective' phase we 'profile' (Volk & Harder 2007:502) our three training corpora, specifying: the number of source language repetitions (within the corpora, and between the corpora and the test data); the size of the corpus; and the homogeneity of the corpus (independent variables). In addition we elicit human judgements on the quality of the corpus data and on the reusability in new contexts of the target language subtitles. During the second 'retrospective' phase we establish the intelligibility and acceptability (dependent variables) of EBMT-produced subtitles through end-user evaluation sessions. 44 native German-speaking subjects took part in this end-user evaluation, and their opinions on the quality of automatically translated subtitles on six video clips taken from the first *Harry Potter* movie, *Harry Potter and the Philosopher's Stone* (2001), were elicited using an interview questionnaire. During the data analysis phase we focus on the relationship between the dependent and independent variables. In addition to conducting human evaluation, we generate automatic metric scores and discuss the relationship of these scores to human judgements on the quality of the subtitles.

In addition to the two main research questions, we investigate some subsidiary questions:

**RQ3**: If the viewer understands the (source language) soundtrack, are they more accepting or less accepting of the EBMT subtitles?

**RQ4**: If the viewer has a 'linguistic background', are they more accepting or less accepting of the EBMT subtitles?

**RQ5:** If the viewer has prior knowledge of the movie or related material such as books, are they more accepting or less accepting of the EBMT subtitles?

As is clear from the research questions above, the quality criteria we measure in the current research are intelligibility and acceptability. We further divide intelligibility into comprehensibility and readability, and assume that intelligibility is a necessary although not sufficient condition for acceptability. Acceptability, in turn, also depends on the characteristics of style and well-formedness. In this research *comprehensibility* is understood as the extent to which a text is easy to understand (Halliday in Van Slype 1979:62). *Readability* is defined as the extent to which a text can be read and understood in a prescribed time-frame (cf. Klare 1977 cited in Cadwell 2008:12). The definitions of comprehensibility and readability overlap to some extent, and combining judgements on comprehensibility with judgements on readability of the subtitles gives us a sound measurement of their intelligibility.

*Style* is defined in this study as the extent to which the translation uses the language appropriate to its content and intention (Hutchins & Somers 1992:163). It should be noted that style is distinct from readability, as a text may be highly readable but in an inappropriate style (FEMTI).

Lastly *well-formedness* is defined as the degree to which the output respects the reference rules of the target language at the specified linguistic level (Flanagan 1994, Arnold et al. 2003 and Loffler-Laurian 1983). Combining two criteria such as style and well-formedness provides a good insight into human judgements of acceptability, a concept that is otherwise difficult to define.

According to Hutchins & Somers (1992:163) the three most obvious tests of translation quality are: fidelity, intelligibility and style. We have already discussed intelligibility and style, but have not yet mentioned *fidelity* or *accuracy*. For the current study we did

not explicitly gather human judgements on the accuracy of the machine-generated subtitles. The context of this evaluation differs from other MT human evaluations given the multimodal environment. While the subjects evaluate the subtitles, their judgements on intelligibility and acceptability are influenced not just by the subtitles themselves, but also by factors such as the timing of the subtitles or the synchronisation between the verbal and visual elements on the screen. While viewers of subtitles have extra semiotic channels available to them (compared to users of other kinds of texts), one thing they do not normally have available to them is the source text (script) on which basis target language subtitles were translated. Even if viewers understand the soundtrack language in a subtitled movie, for example, this does not mean that they can judge the 'accuracy' of the subtitles.

### Significance of the *Harry Potter* movie clips

For this study we use clips taken from the first *Harry Potter* movie, *Harry Potter and the Philosopher's Stone* (2001). This choice was motivated by the fact that this title is one of a series of *Harry Potter* movies. Movies within the same series belong to the same genre, and many of the main characters are the same throughout. It was thought possible that phrases used in the first movie might also be used in subsequent movies. Given that EBMT technology works on the basis of recycling translations for recurring source-language segments, it was considered worthwhile investigating if this technology would aid the subtitling of subsequent movies in the same series.

### Structure of the Thesis

Chapter 1 presents a review of the relevant literature on text recycling and intertextuality, and subtitling as a form of translation. The role technology currently plays in subtitling is outlined, and a brief sketch of contemporary Machine Translation (MT) is drawn. In particular we describe the Example-Based MT (EBMT) system which is employed in this study. The chapter concludes with an overview of previous related studies in automated subtitling, which form a basis for the current study.

Chapter 2 presents a review of Machine Translation evaluation, one of the main topics of interest in this study. It describes the different approaches to human MT evaluation, and the apparent move away from human approaches to automatic evaluation methods. We situate our human evaluation among previous research, outlining differences

between our human-based approach and previous research. We also describe our use of the Framework for the Evaluation of Machine Translation in ISLE (FEMTI) when defining our quality characteristics and measurement techniques.

Chapter 3 presents the methodology used in this study. The chapter begins by restating our research questions and goes on to describe the research design adopted in this study. The Example-Based Machine Translation (EBMT) system used is described, and theoretical and practical research design issues are discussed. Finally, the two-phase approach we adopt in the study, which includes a prospective and a retrospective phase, is explained.

Chapter 4 presents the results from the prospective phase of the study. Firstly we analyse the results from the corpus-analysis stage, and comment on the corpus profiles. Then we analyse the results from the human evaluation of the reusability of TL subtitles, and relate these results to the corpus profiling stage.

Chapter 5 presents the results from the retrospective phase of the study, in which we analyse questionnaire data. We investigate if any relationships exist between the training corpus profiles established in Chapter 4 and viewers' judgements on the intelligibility and acceptability of EBMT-generated subtitles discussed in this Chapter. We also investigate whether other variables (e.g. subjects' knowledge of the soundtrack language) impact on their judgements.

Chapter 6 presents a discussion of the results. It reflects on the research questions set out at the beginning of the study and the methodology used to address them. We ask whether the objectives of the study have been met, and also highlight unexpected findings. We identify areas of future research leading on from this study, and possible improvements that could be incorporated into subsequent studies.

# - Chapter 1
# - Literature Review

# 1 Literature Review

This chapter introduces the many intertwined concepts pertinent to the study. We begin by looking at the definition of a text, textuality and the semiotic notion of intertextuality (1.1). This leads us onto a discussion of Audiovisual Translation (AVT) (1.2), the domain in which we introduced an automated approach. During the discussion on this discipline, we incorporate references to text, intertextuality, originality and digitalisation, in addition to subtitling, the method of AVT we employ in this work. Following this we introduce the broad topic of translation technology (1.3). This sets the scene for a short history of Machine Translation (MT) and MT systems (1.4). Within this section we mention the various approaches to MT, and give a detailed description of Corpus-Based MT approaches. This is followed by comments on two topics which are often ignored in the literature, namely repetition and reusability. Then we refer to previous studies that relate to the current one, and highlight differences between the approaches taken within these studies and the evaluation strategies employed here (1.5). One of these studies is MovieTrans: Rapid, Memory-Based Audiovisual Translations (MovRat), a direct precursor to the current study, and one to which the present researcher contributed.

## 1.1 Text and Intertextuality

The term *text* is derived from the Latin *texere*, *textum* meaning 'to weave', 'woven'. This suggests that a text has many threads, which are interwoven with each other. A text is not a stand-alone piece of writing but what Barthes (1977:146-7) describes as "a multidimensional space in which a variety of writings, none of them original, blend and clash." In a semiotic sense, the signs within a text can take the form of words, images, sounds and/or gestures, which are associated with a genre and a particular medium of communication (Allen 2000:1). Chandler (2007:253) comments that 'medium' is interpreted in a variety of ways by different theorists, some using broad categories such as speech and writing or broadcasting, or relating the term to specific technical forms within the mass media, including radio, television, newspapers, books, photographs, films and records, as well as the media of interpersonal communication, including letter, fax, e-mail, video-conferencing and computer-based chat systems. The medium that this research focuses on is interlingual subtitling. For this research, we are concerned with

the relationship between speech (soundtrack), image and the written (subtitles), and consider text to cover signs relating to these three channels of communication.

Theorists such as de Beaugrande & Dressler (1981) define text as a communicative occurrence that is characterised by a quality known as *textuality*. According to de Beaugrande & Dressler (ibid:3), there are seven standards of textuality: *cohesion*, *coherence*, *intentionality*, *acceptability*, *informativity*, *situationality*, and *intertextuality*. Two of these standards are of particular interest to this study, namely acceptability and intertextuality. We will come back to acceptability in the next chapter when we discuss our evaluation approach. Intertextuality is discussed below.

### 1.1.1 Intertextuality

According to Allen (2000:1) *intertextuality* is a term used to describe texts which are lacking in any kind of independent meaning. Other definitions in the literature describe it as the reference of one text to another; the idea that texts are made up of other texts (Genette 1997); and the "reuse of existing written sources in the creation of a new text" (Clough et al. 2002:1). On the basis of these definitions, we can say that a text is dependent on other texts, and those texts in turn are dependent on a different set of texts. No text is an entirely original piece of work; each text producer has read many texts and is influenced by those they have previously read. Reading existing texts influences what we write in the future. These definitions deal with both the reuse of the texts, and the meaning relations between texts. The reuse of textual material in another text disrupts the conventional 'linearity' of texts. It also has implications for how we understand concepts such as authorship and plagiarism, with Chandler (2002) reminding us that reuse of others' texts was expected in the middle ages. He goes on to quote Goldschmidt saying "before 1500 or thereabouts people did not attach the same importance to ascertaining the precise identity of the author of a book they were reading or quoting as we do now" (Goldschmidt 1943:88 cited in Chandler 2002).

Roland Barthes, who has been described as one of the most "articulate of all writers on the concept of intertextuality" (Allen ibid:61), argued that a text is a plurality of voices, words, utterances and other texts. If it were possible to see inside the mind of an author, we would not find original thought, but rather what he describes as the 'already-read' and 'already-written'. The meaning of a text does not come from the author, but rather

from the intertextual nature of the language. This idea is brought to life in his 1968 essay 'Death of an Author', which puts forward his theory of intertextuality.

Whatever way we intend to interpret intertextuality, the underlying message of this term seems to be that it "promotes a new vision of meaning, and thus of authorship and reading: a vision resistant to ingrained notions of originality, uniqueness, singularity and autonomy" (Allen ibid:6).

*Transtextuality* is a term proposed by structuralist theorist Gerard Genette. He built on work of Bakhtin (1981, 1986) and Kristeva (1980a, 1980b), and his idea refers to "all that sets the text in a relationship, whether obvious or concealed, with other texts" (Genette 1997:1). Within the term transtextuality, there are five types of transtextual relations, *intertextuality* being one of these. Genette uses the term to refer to the "effective co-presence of two texts" and "the actual presence of one text within another" (ibid:1-2).

## 1.2 Audiovisual Translation

Audiovisual Translation (AVT) was once considered to fall outside the scope of Translation Studies (TS), but over the past decade or so, due to factors such as the centenary of the cinema in 1995, language minorities and the growth of technology, AVT has emerged as "at least a sub-domain of TS" (Gambier 2008:13). Most notably the number of theses and studies in AVT has also grown steadily during this time. However, most of this research is from a linguistic perspective, even though AVT is "a multisemiotic blend of many different elements such as images, sounds, language (oral and written), colours, proxemics and gestures" (ibid:11). Subtitling has been referred to at various stages as *constrained translation* and *a necessary evil (*Marleau 1982, Titford 1982 and Mayoral et al. 1988), however, all over Europe AVT "plays an increasingly important role in modern mass communication" (Gottlieb 1992:161).

AVT covers various types of translation, including dubbing, subtitling, surtitling, voice-over, interpreting and audio description and can be described as "the academic field which studies the new reality of a society which is media-oriented" (Orero 2004:vii-xiii). All types of AVT are growing in popularity given the growth in new technology and language awareness, as well as changing language policies. Gambier (2003:172)

notes that the various types of AVT can be categorised into two main groups: dominant and challenging. The dominant types include interlingual subtitling and dubbing, probably the two best known types of 'screen translation'. Challenging types include intralingual subtitling, live or real-time subtitling and audio description (AD). In this study we focus on a dominant type of AVT, interlingual subtitling. First of all, however, we turn our attention to subtitling in general, define the concept, and outline the main differences between intra- and interlingual subtitling. This is followed by a discussion of the main uses of interlingual subtitles.

### 1.2.1 Subtitling

> Subtitles are not a replacement of anything, but an addition to a film – they form an overlay, so that one has a kind of simultaneous bicultural interpretation of what is going on.
>
> (Hofstadter 1997 cited in Ivarsson & Carroll 1998:iii)

Subtitling can be split up into *intralingual subtitling* or *closed captions* (also known as subtitling for the deaf and the hard-of-hearing (SDH)), and *interlingual subtitling* or *open captions* (the main audience of these subtitles is those whose mother tongue is not that of the original soundtrack) (Gambier 2003:174). The difference between the two types of subtitling relates to the different requirements of hard-of-hearing and hearing viewers. The current research focuses only on subtitling for hearing viewers. Interlingual subtitling involves translating from the oral dialogue to one to two written lines of text, moving from one language into another (or possibly two in the case of bilingual countries, e.g. Belgium). Given that many viewers use subtitles to understand a movie, for example, all semiotic information in the original dialogue should be recoverable from the subtitles. But subtitling is not simply regarded as a linguistic process; the effectiveness of subtitles is crucially dependent upon semiotic relations between the linguistic and visual content (de Linde & Kay 1999:74).

Two differences often mentioned between the 'traditional' translation of texts and the translation of subtitles are time and space constraints, resulting in a reduction in verbal content and perhaps a certain 'loss' of information. According to Luyken et al. (1991), the amount of time a subtitle remains on the screen depends on three factors:

- The amount of text
- The average reading speed of the viewers
- The constant minimum interval between subtitles

In general the minimum time for a short subtitle is at least one and a half seconds, with the maximum time for a full two-line subtitle not exceeding five to six seconds. A short subtitle has to remain on the screen for this minimum time to avoid the risk of the viewer missing the subtitle and to avoid a 'flashing' effect which is disruptive to the viewer. There also has to be a minimal pause between longer subtitles so that the eye can register that a new subtitle has appeared. The reason for these time constraints is explained very simply by Ivarsson & Carroll (1998:64) when they say "a lot more than subtitles meets the eye." When viewers are watching a subtitled movie, they are not simply reading the subtitles, but also listening to the background sounds, the soundtrack and looking at the image in parallel (Ivarsson & Carroll 1998, Gambier 2003).

Space constraints associated with subtitling refer to the number of characters permitted in a one or two-line subtitle. This is restricted due to the size of the screen on which the subtitles appear. Subtitle norms have been established and it is suggested that two-line subtitles contain up to 80 characters and remain on the screen for between five and six seconds. This results in an average reading speed of 175 words per minute (Ivarsson & Carroll ibid:67). In addition to these two main technical considerations, the display and format of the subtitles should be considered. This includes the position of the subtitles on the screen: this often depends on cultural preferences and two common positions are centred at the bottom of the screen (Europe, the US and Japan[1]); and in a vertical column to the right or left (Korea).

In our methodology chapter we describe in detail how we put the EBMT subtitles onto the movie clips used in the evaluation sessions. At this point it suffices to say we followed suggested guidelines (Ivarsson & Carroll 1998, Cerón 2001) for timing the subtitles, introducing line breaks to make two-line subtitles, and correctly synchronising the subtitle and image. We followed the European convention of positioning the subtitles in the centre at the bottom of the screen. The subtitles were generated using the

[1] In Japan, the primary subtitles are no longer shown vertically on the screen.

default font size and colour offered by the subtitling software used for the study (Subtitle Workshop), white lettering with black outline.[2]

Time and space constraints have always been a defining factor of subtitles. We have seen in the literature that subtitles were once described as a 'constrained translation'. In much of the earlier AVT literature, interlingual subtitles were stereotyped as being 'inferior' to the original dialogue. Over the years there has been literature published that opposes this restrictive view, with Reid (1978 cited in Gambier 2008:16) defending subtitling as 'the intelligent solution', Gottlieb (1994) describing subtitling as a 'diagonal translation' and Gambier (2006 cited in Gambier 2008) speaking of "'selective translation', in which any translation is necessarily 'constrained' by the medium concerned, whether it be a comic strip, an illustrated book, a children's book, a scientific paper or an exhibition catalogue." It is clear from watching a subtitled programme that the subtitles are not a verbatim representation of the spoken dialogue. Nevertheless, is the viewer correct in thinking that subtitling implies reduction, resulting in a 'loss' of information? Gottlieb (2005:19) addresses this question of reduction in verbal content and argues that text reduction is neither semiotically nor technically motivated, nor does it have anything to do with reading speeds, but rather with producing a good quality translation. He highlights how subtitling practices have changed over time: it has been assumed that today's viewers of subtitles are probably faster readers than previous generations. Becquemont (1996:147 cited in Díaz Cintas & Remael 2007:98) mentions that the change in subtitling practices is evident in France, where subtitles were kept on screen for up to six seconds during the 1980s, and during the 1990s this was shortened to five or four and a half seconds, as six seconds was considered excessive by some professionals. Furthermore, Gottlieb (ibid) points out that commercial TV stations and some within the DVD industry have presupposed this change in reading speeds, as previously 12 subtitle characters per second (cps) were displayed, and this number has been raised to 16 cps. Gottlieb believes that this should mean that there is no need for the usual reduction of 20-40% of semantic and stylistic content of spoken dialogue, given the longer subtitles.

[2] For future projects that ask end-users to make judgements on automatically-generated subtitles, we would suggest using a professional subtitler and a professional subtitling package (e.g., Swift) in order to control the cosmetic aspect of the subtitles (e.g., font, subtitle position) and to eliminate possible 'noise' contributing to the viewer perception of the subtitles used in the human evaluations.

We would agree with this point regarding subtitle speeds; however, we would also draw on Ivarsson & Carroll's (1998) point regarding the reading of subtitles being a process that is integrated with listening to the soundtrack and watching the image. Increasing subtitle length does not necessarily increase quality and text omission may simply be a strategy to remove some redundant utterances, which is related to Gottlieb's point about translation quality. He comments that written text as a language mode is more concise than oral discourse, and therefore for translation quality purposes, subtitles can be condensed. This is a very valid point, given that subtitling occurs in a polysemiotic[3] context, and what might be missing in the subtitles can be filled by the other semiotic channels. Gottlieb's (2005) main observation is that time and space constraints of subtitling can be used as an excuse to 'normalise' the text, removing any elements which might seem troublesome to translate or perhaps controversial in the target language.

Subtitles are distributed widely: they are available on almost all TV programmes and video media, including news programmes, documentaries, quiz shows, soaps, films on TV, video cassettes (VHS) and digital versatile disk (DVD). At one stage subtitling might have been described as a 'minor' art form (Ivarsson & Carroll 1998:5); however, it is certainly not a minor activity in many countries given the amount of subtitling performed annually (cf. Luyken et al. 1991). There are countries typically termed 'subtitling countries' and 'dubbing countries'. Examples of subtitling countries in Europe are Scandinavia, Ireland, Portugal and Greece; and examples of dubbing countries include Germany, France and Italy. Historically, dubbing was favoured in countries with strong nationalistic behaviours, defending the national language, and on occasions banning any subtitles from appearing on TV or cinema programmes (ibid:10). Since then times have changed and Ivarsson & Carroll (ibid:v) point out that:

> Subtitling will not remain the domain of a few 'subtitling countries'. […] countries which traditionally used to dub films are turning to subtitles for cost reasons on the one hand, but also because of changing audience demands.

[3] Chaume (2004:16 cited in Gottlieb 2005:2) defines polysemiotic as "a semiotic construct comprising several signifying codes that operate simultaneously in the production of meaning."

One of main advantages of subtitling over dubbing is the cost factor: dubbing is ten to twenty times more expensive than subtitling (ibid:36). There are, of course, advantages and disadvantages with both dubbing and subtitling (cf. Ivarsson 1992, Ivarsson & Carroll 1998, Koolstra et al. 2002, Choi 2003). Advantages of subtitling include low costs, and being able to hear the original soundtrack (and subtitling does not suffer from the problem whereby the dubbing actor assigned to a particular Hollywood actor is reassigned to a new Hollywood actor, much to the dismay of the audience). Subtitling does not interfere with gestures, body language and facial expressions on screen, which when combined together are a significant source of information for viewers, and subtitles are invaluable in the learning of a foreign language. Disadvantages of subtitling include interference with the image, the fact that viewers' attention is diverted from the image when they read the subtitles, that there may be mistakes in the subtitles (both grammatical and typos), and that sometimes subtitles are not properly synchronised with the image, either appearing on the screen too early/too late, or being removed too early/too late (Ivarsson & Carroll ibid:35). Another factor is that providing subtitles on AV material assumes that the viewer is literate (e.g. children's shows are typically dubbed for this reason). Conversely, advantages of dubbing a movie or programme mean that no texts are projected over the image, maintaining the unity of picture and sound, and viewers of dubbed AV material regard the material as being familiar, since they hear a language which they understand (Koolstra et al. 2002:339), which allows them to become totally immersed in the dubbed programme they are watching. Koolstra et al. (ibid) outline a study (Huysmans & de Haan 2001) which showed that if listening to a dubbed TV programme is a secondary activity to say a primary activity such as reading a newspaper, the listening process is easily achieved, even if it is with only 'half an ear'. They conclude that 'only' having to listen to the soundtrack is less mentally demanding than having to read the subtitles (op. cit.:335). One of the major disadvantages of dubbing is that because the original soundtrack is totally removed and replaced by a foreign language soundtrack, dialogues can be adapted easily, which makes dubbing scenarios more vulnerable to manipulation and censorship than subtitling scenarios. Ivarsson & Carroll (1998:108) point out that:

> A person who reads a book in translation or sees a dubbed film must go to the original text to check what they suspect is a faulty translation, and very few people take this trouble...the subtitler is in a much more vulnerable position, since the original is available for all to see and hear.

Díaz Cintas & Remael (2007:57) also highlight how "subtitles must stand up to the scrutiny of an audience that may have some knowledge of the original language" and call subtitling an instance of 'vulnerable translation'.

These arguments cast some doubt on Luyken et al's. (1991:73, emphasis in original) contention that dubbing is (necessarily) "the replacement of the original speech by a voice-track which is a *faithful translation* of the original speech." We comment on this point again in Chapter 5.

Despite the reported disadvantages of subtitling, in a society where the cost of technologies and services is decreasing, subtitling is still a very good option that allows increasingly linguistically diverse populations access to AV texts. Digitalisation has had and continues to have a significant impact on AVT in relation to subtitles, production of images and sounds, setting, costumes, filming and editing, distribution and screening (Gambier 2008:25). With digitalisation came the idea of introducing language technologies into the subtitling process, a point which is further explored in section 1.5 below.

The German-speaking countries, including Germany, Austria and Switzerland, are traditionally considered to be 'dubbing countries'. In these countries all TV programmes that do not have German as the original language soundtrack are dubbed (including films, chat shows, series and children's cartoons). The result of this for many years has been that deaf or hard-of-hearing (HOH) viewers have had difficulties in accessing the material shown on TV, and hearing viewers have not been able to listen to the original soundtrack of foreign language movies. With the arrival of DVD, however, this situation has changed: a growing number of movies available now on DVD contain intralingual subtitles (German subtitles on a German soundtrack movie for deaf or HOH viewers), and pressure groups in Germany have managed to get many foreign language movies marketed with two interlingual subtitle tracks – one for hearing viewers and one adapted for deaf or HOH viewers. This has opened up subtitling opportunities in

countries like Germany, for example, where once only dubbed movies could be viewed; as already mentioned, DVD has now introduced interactivity and the ability for viewers to select their preferred mode of language support (O'Hagan 2007:157). However, given the fact these new subtitle features are available only on DVD, many viewers in Germany are still not exposed to subtitles, and are used to watching only dubbed AV material.

Luyken et al. (1991) look at various 'dubbing' and 'subtitling' countries, concentrating on countries within Europe, and find that the preference of television audiences for the type of language transfer used on TV programmes is determined by "their familiarity with, and conditioning to a specific method" (ibid:185). Focusing on Germany in this instance, 78% of Germans[4] surveyed favoured watching the programmes in dubbed form. It could be asked if the majority of viewers are satisfied watching dubbed programmes, why would the TV stations start introducing subtitles, and perhaps lose some viewers? However, Luyken et al. (ibid) did notice that many of the countries who usually stayed with one type of language transfer in the past, in most cases dubbing, seemed to be opening up to the idea of watching foreign language films with another type of language transfer (usually subtitles), with results from The Netherlands over a 13 year period to back up these views. Interestingly enough, some of the viewers who were open to change were among the most critical viewing groups: the less well-educated (a 20% increase of preference for subtitled programmes) and the elderly (a 16% increase of preference for subtitled programmes). Television reading speeds for both of these groups might be expected to be lower than those belonging to the well-educated and young people categories. The study also shows that audiences become very accustomed to one particular type of audiovisual transfer; this can mean that audience research merely reflects the viewers' attitudes and behaviour towards what they are being offered, and not towards the types of audiovisual language transfer available (ibid:188). In relation to the current study, if the subjects have negative attitudes towards viewing subtitles prior to the evaluation, their judgements could negatively affect the results. This point is explored further in section 5.2.2.

[4] It must be noted here that at the time of this study Germany was not fully unified, and therefore the study relates to figures based only on the former West Germany.

### 1.2.2 Subtitling and Technology

> Audiovisual translation (AVT) in general, and in subtitling in particular, has an umbilical relationship with technology, which to a large degree determines it.
>
> (Díaz Cintas 2005:1)

One of the original methods to include subtitles in a film was to project the subtitles onto the screen beside or below the intertitles,[5] a task which was very strenuous for the operator, resulting in the technique never becoming a huge success (Ivarsson & Carroll 1998:9). The problems with this technique were related to synchronising the film and the projected subtitles. With advances in technology, it became easier to project subtitles, and nowadays subtitlers have the benefit of time codes, which can control the appearance and disappearance of subtitles precisely. Time codes were available to producers of subtitles on cinema film long before they were made available for TV/Video/DVD production. Their use with these types of media has brought about radical changes in the cueing of subtitles. A time code is a type of address that marks each individual frame of a videotape, allowing a subtitler to easily identify every frame of the film. This address is an 8-digit number, indicating hours:minutes:seconds.frames, e.g., 12:38:32.06.

The arrival of the computer within the world of translation has greatly changed work practices, and probably even more so within the field of subtitling. There are now many computer programs specifically designed for subtitling, allowing subtitlers to get rid of the once popular subtitling program that required a computer, video player and an external television monitor in order to carry out their work. Instead, subtitlers can now carry out their work using a computer, subtitling software, and a digitized copy of the audiovisual programme they want to subtitle (Díaz Cintas ibid:2).

Experienced subtitlers nowadays have dedicated subtitling software applications in which to generate subtitles. A word processing program specially designed for subtitling is installed on the computer, and displays the subtitles on the computer screen as they will appear on the television or movie screen. The workstation also includes a spellcheck program. Given the role technology plays in the domain of subtitling

[5] Intertitles were the result of the first efforts made to convey dialogue of actors to the audience, first seen in 1903; they are text, drawn or printed on paper or cardboard, filmed and inserted between sequences of the film. They became known as subtitles a while later (Ivarsson and Carroll 1998:9).

nowadays, the profile of subtitlers has also changed, meaning they must have good technical knowledge, in addition to linguistic competence and socio-cultural and subject knowledge. Ivarsson & Carroll (1998:150) point out that translation tools such as translation memory (TM) tools and machine translation (MT) systems could well be an integral part of a subtitler's workstation in the future, helping them to speed up the subtitling process, removing the tedious, repetitive work, and helping to standardise terminology, but Díaz Cintas (ibid:2) has argued that the value of TM tools in AVT "is questionable and still to be researched." He goes on to claim that while corpus studies have benefitted other areas of translation, they do not yet appear to have made their entry into the field of AVT. He also describes the use of machine-assisted translation as an 'incipient reality' in the USA, a country with a long history of SDH.

Taylor (2006) focuses on repetitive language in subtitling, arguing that it could be possible to 'predict' what the actors will say next. Based on genre analysis research (cf. Swales 1990, Bhatia 1993, Ventola & Mauranen 1996), Taylor talks specifically about how language can be categorised into different subdivisions, including genres, subgenres and his own term 'genrelets' (ibid:4), and maintains that the language of film is an entity of its own compared to general, everyday spoken language. Given these two premises, he believes it could be possible to introduce strategies and techniques familiar from TM technology, to the subtitling process.

An important and fast growing area of the film industry is DVD releases. Since first appearing on the market in 1997, the production and sales of DVDs have surpassed anybody's expectations. The introduction of the DVD has also brought to the forefront the use of subtitling with film releases. Audio and video material as well as all types of electronic documents can be recorded onto a DVD. A DVD can store from 4 to 28 times as much data as a normal compact disk (CD), allowing for DVDs to store films subtitled in many different languages – in fact, a DVD can contain up to 8 different dubbed versions of the same programme, and up to 32 subtitle tracks in different languages (Armstrong et al. 2006c). This has meant that TV series which were previously only broadcast dubbed on Spanish, Italian or German television, are now sold on DVD with both a dubbed and subtitled version (examples include *Friends*, *Sex and the City* and *The Simpsons*). Since the production of subtitles requires only a small investment, it is

no longer the case that only 'art house' films are being released on DVD with subtitled versions (Díaz Cintas 2005:3).

Like Díaz Cintas (2005), Gambier (2008:23) comments on the role technology now plays in subtitling. He outlines changes in subtitling practices, giving an example of how introducing emoticons, pictograms and abbreviations, very similar to the formats used in text messaging, into subtitling was once thought very outlandish, but that there are two particular cases where a shift in language strategies is visible: the introduction of emoticons to depict different moods in the HOH subtitles for a Portuguese commercial TV station; and the use of spelling changes to promote accessibility, as evidenced by the City of Montréal's online homepage, where "phonetic" spellings replace "correct spellings". These changes are related to the new digital era in subtitling, and based on these changes, Gambier (ibid:24) argues that perhaps automatically translated output should be viewed differently to human-generated subtitles, as this automatic output could satisfy some users not requiring "a polished, finely honed text."

### 1.2.3 Text, Intertextuality, Originality and Translation Strategies

Traditional translation studies deals with texts that are considered 'verbal only', and therefore communicate through one semiotic channel only. However, this monosemiotic label is slightly misleading as "no text can be made entirely of verbal signs because such signs always need some sort of physical support" (Zabalbeascoa 1997:338 cited in Gottlieb ibid:2). Gottlieb adds that the 'physical support' noted by Zabalbeascoa "gains semantic momentum in genuinely polysemiotic texts" (ibid), for example AV texts. Gambier (2008:22) makes a distinction between traditional translation texts and AV texts explaining that AV texts are "short-lived and do not fit readily into the traditional dichotomy between source and target text." He also points out that these texts will not be read over and over for decades or stored in archives for frequent use and reference. We should comment here that this definition is changing with the emergence of DVD and Internet sites such as YouTube.[6] Another issue is the fact that subtitles are readily extractable (albeit illegally). Gambier, like Gottlieb (2005), regards the multimodality[7] of an AV text as perhaps their distinguishing feature.

[6] <http:// www.youtube.com> [Accessed 27 July 2009].

[7] According to Kress & Leeuwen (2001:2) a 'multimodal' text is one that uses several modes of communication (e.g., speech, image, writing) in an integrated way to transmit a message. They argue that multimodality is a characteristic of many kinds of text in the modern world (resulting from the impact of

Moreover, Gambier (ibid) questions whether a multimodal distinction really exists between texts used in TS and AVT, asking whether all texts could be deemed multimodal, giving examples such as tourist brochures, children's literature and instruction leaflets, to name but a few, that use a combination of text and images. This view seems to be shared by the multimodal discourse analysis community (e.g., O'Halloran 2004). Gambier poses the question of whether text means the same thing in literary translation, conference interpreting and AVT, and whether we can continue to speak of texts as a linear arrangement of sentences (ibid). The intertextual characteristic of texts compels us to view texts in a different light and to perhaps review our concept of language norms, for instance. Subtitles have been likened to literary texts in a translation context (Volk 2008), and yet Gambier (2003:178) points out that AV translators are sometimes grouped with literary translators, sometimes with interpreters, and sometimes with technical translators. This shows that there is no clear-cut category to which AV texts, or subtitlers, belong.

Originality is a problematic notion in textual studies in general, and this is no less true of translated AV texts. Gottlieb (2005:23) argues that "it is probably no exaggeration to say that there exists no form of translation in which the notion of an 'original version' is completely sustainable" and goes on to say that the notion of originality not only applies to language ("Which is the original language?"), but also to semiotics ("Which version should be considered the original?"). The problematic nature of the source text is also commented on by Gambier, who asks whether a movie subtitler translates from the script or directly from the soundtrack (Gambier 2004). Gottlieb believes the script represents the original (writer's) intention, as the dialogue is written to be spoken. However, the final recording of a movie is not usually based on the script that was first presented and a subtitler should translate from the dialogue, a practice which is not carried out very often (Gambier 2004, Gottlieb 2005). This is an example of subtitlers working from a pre-production vs. a post-production dialogue list. A dialogue list is "the compilation of all the dialogue exchanges uttered in the film and it is a document usually supplied by the film distributor or producer of the film" (Díaz Cintas & Remael 2007:74). Subtitlers normally work with a dialogue list of some kind except for bonus

digitalisation). *Multimodal* and *polysemiotic* are terms that are used synonymously here, with multimodal representing a more 'theory neutral' concept.

materials which often lack such lists. In some cases dialogue lists are used for reference (in cases where subtitlers are allowed to do their own cueing), but in other cases the dialogue on the list is already compressed into subtitles in the original language and subtitlers must adhere to the cueing of these 'master' subtitles (ibid:75).

The final concept we discuss is translation strategies of translators. Gottlieb (2005:21) puts forward two counter-arguments to concepts commonly accepted in translation studies circles: first, translators do not often make conscious choices when implementing translation strategies; and second, translators often see only one solution. He describes translation strategies as "the guiding principle behind all translational activity" (ibid), which is further supported by Zabalbeascoa (1997:337 cited in Gottlieb ibid):

> Each part or aspect of a translation can be perceived as the outcome of a process of choosing among various possible solutions in the light of all the operative factors of the moment.

However, as Gottlieb (op. cit.) highlights, when even talented translators are working under time pressure, the "process of choosing among various possible solutions" simply does not take place. Time constraints are an integral part of the subtitling process, both in relation to 'spotting' in and out times of subtitles on a screen and the turnaround time between a movie being shown in a cinema and the release of the DVD version with subtitles (O'Hagan 2007). Eco (2004:182 cited in Gottlieb ibid) comments that:

> Translators simply behave like polyglots, because in some way they already know that in the target language a given thing is expressed so and so. They follow their instinct, as does every fluent bilingual person.

Since the move from analogue to the more compact digital technology in the 1990s, time pressure on subtitlers has grown considerably. This has meant that the conventional translation process has had to change, both in general and in terms of subtitling, and perhaps resulting in subtitlers not being able to consider a variety of possible translations before creating a subtitle. They might also use the same subtitle solution in several places throughout a programme/movie, if the subtitle meets the time and space constraints, and it is suitable in the given context.

The discussion presented here has proposed that texts are intertextual in nature, and that this is especially true of polysemiotic texts, such as those used in subtitling. It has also problematised the notion of originality, and highlighted constraints placed on subtitlers' creativity, especially due to increased time pressure. The introduction of automated technologies into the subtitling domain is motivated by three constraints: increasingly tight deadlines imposed on subtitlers for the distribution of varied material; the reduction in costs of subtitling (including a reduction in rates of pay to subtitlers); and the higher volume of material to be subtitled, due to the arrival of digital TV and DVD (O'Hagan 2003a, 2003b, Gambier 2008, Armstrong et al. 2006c). Given that the concept of reusing texts is not foreign to AVT practice, it is a natural progression to introduce a technology, such as Example-Based Machine Translation, that is also based on the idea of reusing text, and whose aim is to assist the subtitler at a time when digitalisation has, and will continue to have, an increasingly marked effect on the role of subtitlers. Such considerations have led to an examination of whether technology can respond to the subtitler's needs.

## 1.3 Translation Technology

> The extent to which one can automate translation is an indication of the extent to which one can automate 'thinking'.
>
> (Arnold et al. 1994:5)

Translation technology is a general term used to describe the technologies or computerized tools available to translators to help them do their job. As Bowker (2002:6) mentions, this can include well-known tools and resources such as word processors, grammar checkers, email, and the World Wide Web (WWW), and less well-known technologies such as terminology-management systems, bilingual concordancers and corpus-analysis tools. Two of the most common technologies are probably Translation Memory (TM) and Machine Translation (MT), which will be discussed in more detail in the following sections. Bowker (ibid:4) outlines the 'traditional' distinction between the two technologies in a clear way when she says MT systems try to replace the translator, while TM tools support translators by helping them work more efficiently. However, when we outline the rationale behind using MT in this research, it will be clear that it is not motivated by a desire to replace the translator, but rather focused on aiding the subtitler in this digital era.

### 1.3.1 Why do we need Translation Technology?

"Translation between human languages has been a need in society for thousands of years" (Trujillo 1999:x). This need is steadily growing resulting from globalisation, given that if we want to sell our products to foreign markets, we need to provide the appropriate information in the language of the target market with these products.

Many people have asked over the years why it is necessary to fund an area of research like MT, when it will never be capable of taking over the position of the human translator. This relates to the misconceptions of what MT is and what it is supposed to do. It is not envisaged that MT will take over fully from human translators in all areas of translation. However, MT is currently capable of carrying out much of the repetitive work and many mundane aspects of translation, allowing human translators to spend more time on the 'creative' side of translation. There is also a need for more 'trivial' translations of disposable texts (e.g. web pages). These can be created on the fly and at no cost to the translator. MT systems are designed for a particular use, and do not claim to be able to translate every piece of natural language produced. There are certain texts, for example, literary works, which an MT system cannot translate. This does not mean that MT is useless – it simply means that the system does not have anthropomorphic characteristics required to translate literary texts. Although MT is imperfect, it is certainly possible and clearly a reality. It is important to be aware of its strong points together with its shortcomings. As Flanagan (1997:25) points out "whether or not MT can help translators is no longer the only question by which to assess MT's success. It is important to look at other uses of the technology as well."

One of the most important influences technology has had on the area of translation is the speed with which tasks can be carried out. MT can also increase the volume of translation throughput. Technology, it seems, also makes it easier to ensure consistency within a translation, although Bowker (2005) argues, this is not always a reality when working with TMs, and other researchers are not convinced of the positive influence translation technology has brought to translation. Mossop (2006:787-793), for example, poses the question of whether or not computers are doing nothing more than speeding up the writing and research processes, rather than aiding translators with their work. He believes these technologies are developed to serve business purposes, and their introduction has possibly changed the mental process of translation, a concern already

voiced by researchers in the field of TMs (cf. Bédard 2000, Delisle 2006). However, looking at proceedings from translation and language related conferences (LREC, EAMT, ACL and IATIS), the many journal and book publications in the area, and collaborations between industry partners and researchers, it is clear that both technologies are in high demand and are used on a regular basis. The ability to remove tedious work from a translator's workload and to use computers to translate technical materials has certainly increased productivity in a society where languages are at the forefront, and market demands on translation have increased ten-fold (Webb 1998, Benis 1999, 2000, Wallis 2006).

## 1.4 A Short Introduction to Machine Translation

Machine Translation (MT) is the process whereby a computer program translates a text from one natural language into another, and has been widely considered a "tangible goal since the late 1940s, with the advent of the digital computer, the concept of stored program and the promise of large storage devices" (Nirenburg et al. 2003:4). The main focus of MT is to produce high-quality natural language output, translating from one language into another, quickly and cost-effectively. The realisation that fully automatic, high-quality output for all domains was far more difficult a task than was originally anticipated, means that MT systems are increasingly being used in conjunction with other types of electronic translation tools (including TM tools), together with the cooperation of humans. That said, there are stand-alone MT systems, which do not use any human input, but in some of these cases, the input is written in a controlled language[8] (for example the KANT[9] system; see also O'Brien 2006 and Roturier 2006 for more detailed research on controlled language). Many academics have written about the colourful history of MT, including Arnold et al. (1994), Hutchins (1986), Trujillo (1999), Nirenburg et al. (2003). Here we provide just a brief sketch of relevant developments in the area.

During the 1960s there were some doubts voiced as to whether or not MT was viable. MT was recognised as an extremely difficult application and was criticised for failing to produce translations equivalent to those of humans. This led to the publication of the

[8] A controlled language is defined as "an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style" (Huijsen 1998:2 cited in O'Brien 2006:6).
[9] <http://www.lti.cs.cmu.edu/Research/Kant/> [Accessed 10 March 2009].

ALPAC[10] report, which crushed the hopes of MT research groups, concluding that there was no shortage of human translators, and it did not look hopeful that MT systems would produce useful translations of general scientific texts (Arnold et al. 1994:11).

MT has proved to be cost effective within large companies (in cases where the input language is controlled, documentation is repetitive and terminology is standardised) (Hutchins 2005:21). The reason MT often fails at translating a text, is one of the reasons a human might not be able to translate a text: they have no or insufficient knowledge of the source language. With the introduction of Corpus-Based MT systems, which are based on the idea that the system can learn from previous translation examples, MT technology is becoming more prevalent. Improvements in MT over the past 50 years are related to many factors: the systems themselves are much more robust, and cost effective, and produce output within seconds. They are more adaptable to the workplace and it is now possible to integrate them into a translator's workstation. However, we must remember when talking about MT systems, that they are not able to translate *any* text in *any* subject and produce a good translation without the help of a translator (ibid:21). As Melby further notes (1997:29), even though MT systems producing 'indicative translations' is the fastest growing use for MT in places such as the European Union administrative centres, it will certainly not replace human translators; instead it complements them. Shuttleworth (2002:124) points out another use of MT systems that involves the combination of TM tools and MT systems. The advantage of this is that different TM tools offer different possibilities for interfacing with MT engines. He argues that the combination and use of the two technologies can be seen as a way of addressing particular criticisms of TM tools, including their dependence on whole sentence repetition, and of reducing the labour-intensive task of building TM databases. For further information on successful commercial MT systems and on MT improvements see Hutchins (2003b).

The next few paragraphs describe how MT can be divided up based on its different uses, according to two active researchers within the MT community. Firstly, Melby (1997:30, outlined in Table 1.1 overleaf) talks about there being four cases where MT can be used. He distinguishes between two different text types, on the one hand, and two different

[10] ALPAC (Automatic Language Processing Advisory Committee) was established in 1964 by the US Government to assess the progress of human language technology and its prospects.

quality requirements, on the other, with the two kinds of text types and quality requirements at either end of a spectrum.

**Table 1.1**: Text types and quality requirements of translations

| | Text Type | Quality Requirements |
|---|---|---|
| **1** | Controlled domain specific text | High quality output |
| **2** | Controlled domain specific text | Indicative output |
| **3** | Dynamic general text | Indicative output |
| **4** | Dynamic general text | High-quality output |

Melby believes that MT is most appropriate in the first three cases shown in Table 1.1, but the majority of translation falls into the fourth case, which requires human translation.

Secondly, Hutchins (2003b:161) divides up translation into three main groups: dissemination, assimilation, interchange. The first of these, dissemination, requires translation of publishable quality, which is only possible with human input (with or without the use of translation technology); the second, assimilation, is the translation of short-lived documents, not traditionally done by professional translators, and often only for use within large organisations. A good example of this is the use of SYSTRAN (Rule-Based MT system) at the European Commission. We could describe this type as 'inward' translation, for use by the commission staff themselves; and finally, the third group, interchange, covers the role of translation in face-to-face interaction, which was, for example, the motivation behind the Verbmobil project,[11] as well as the translation of traditional or electronic mail. The growth of MT in this area is due to its real-time and online capabilities and low costs, which can offset the effects of lower quality, and therefore users voice little objection to potentially poor quality output.

### 1.4.1 Corpus-Based Machine Translation Systems

Machine Translation Systems can be broadly divided up into two paradigms: rule-based and corpus-based approaches. Traditional rule-based approaches to MT require large-scale grammars and rules, which means developers require extensive linguistic expertise, and a substantial amount of manual labour is needed (Gough 2005:1). This

[11] <http://verbmobil.dfki.de/overview-us.html> [Accessed 10 May 2009].

problem with RBMT systems has come to be known as "the knowledge acquisition bottleneck" (Winiwarter 2007:345). Another problem with RBMT systems identified by MT researchers is the fact that the systems are unable to learn from their mistakes (ibid). Problems such as these prompted developers to approach the task of automatic translation from a new angle, and to come up with a more viable solution. Corpus-Based Machine Translation (CBMT) approaches are the proposed solution to many of the issues associated with RBMT systems. Corpus-based approaches, on the other hand, are not associated with grammars and rules, but instead require an aligned bilingual corpus as a prerequisite (Gough ibid:2). They became popular in the mid-to-late 1980s (Nagao 1984). This was a major turning point in the translation of natural languages by computer. The basic aim of corpus-based approaches to MT (also commonly known as data-driven approaches), is to generate new translations by means of a set of previously saved human (or possibly MT) translated examples. Therefore, a bilingual parallel corpus is a prerequisite for any type of CBMT system. This set or corpus of examples contains potentially reusable translations, which can either be reused in their complete form, or else be broken down by the system into useful fragments, and then recombined to produce new translations. CBMT has proven advantages over RBMT: it overcomes the knowledge acquisition problem of RBMT systems; CBMT systems are also generally more robust than RBMT systems, as RBMT systems contain rules developed by linguists based on incomplete theories, meaning these rules cannot cover every possible linguistic phenomenon. This can lead to a lack of robustness. By automatically inferring rules from a corpus (from actual examples instead of relying on formalisms and theories developed by system developers), corpus-based systems become more robust, and can deal with ill-formed or ungrammatical input sentences. And finally, corpus-based systems try to "learn to transfer knowledge automatically on the basis of the large bilingual corpora" (Winiwarter ibid:345, Knight 1997). This means that a corpus-based system has the ability to append newly derived translations to the corpus during the translation process, which can then be used during the translation of previously unseen input. However, with a rule-based system, the system deals with the input sentences in the exact same way each time, and does not learn from what it generated previously (see Gough 2005:16-17 for further examples).

Within the CBMT paradigm, two main frameworks can be distinguished: Statistical Machine Translation and Example-Based Machine Translation.

### 1.4.2 Statistical Machine Translation (SMT)

SMT was first introduced in 1988 at the Second *TMI* conference at Carnegie Mellon University, when IBM's Peter Brown presented an approach quite unlike anything the audience had ever seen before. It was a 'purely statistical', language-independent approach implementing "a highly developed mathematical theory of probability distribution and probability estimation" (Carl & Way 2003:xix). SMT systems learn a translation model from the required bilingual parallel corpus, and they learn a language model from a monolingual corpus. The translation model establishes a set of target language words or phrases which the system deems will be most helpful for the translation of the input source string. It does this by taking into account source and target word and phrase co-occurrence frequencies, sentence lengths and relative sentence positions of source and target words. In addition to this, the language model tries to assemble the words and phrases generated in the best possible order, to produce the best output string. The language model is trained by determining bigram and trigram frequencies occurring in the training data (Way et al. 2005c:6). Both EBMT and SMT systems integrate word-level alignments. Traditionally, only EBMT systems also integrated phrasal alignments, but in recent years, research in SMT has started to include phrasal alignments, with evidence of improved translation performance since their integration (see Och et al. 1999, Yamada & Knight 2001, Charniak et al. 2003, Koehn et al. 2003). SMT is by far the more dominant approach within the CBMT paradigm. SMT is also being used as the translation engine of commercial translation products, including Language Weaver[12] and the free online MT system Google Translate.[13]

The distinction between the two CBMT approaches is not as clear-cut as it once was, with each approach now implementing functions previously more characteristic of the other. The next section introduces EBMT, and comments on the visible crossover between it and SMT.

[12] <http://www.languageweaver.com/page/home/> [Accessed 10 March 2009].
[13] <http://www.google.com/language_tools?hl=en> [Accessed 10 March 2009].

### 1.4.3 Example-Based Machine Translation (EBMT)

The literature on EBMT has been growing steadily in the last number of years, with research in MT now focussing mainly on data-driven techniques. Somers (1999) gives an excellent review of EBMT and a comprehensive classification of the broad variety of MT research falling within the example-based model. Turcato & Popowich (2003) take Somers' paper as a starting point and try to take further steps in answering questions regarding the classification of EBMT systems. They stress that it is not enough to define EBMT as simply MT which makes use of databases of translation examples, rather what is important is how the data are used in translation operations. Carl & Way (ibid) present a comprehensive overview of the recent advances in EBMT, which notes the historical, technological and philosophical background of the approach. More recently Hutchins' (2006) paper reviews Carl & Way's published collection, surveying the basic processes, methods, main problems and tasks of EBMT, building on work already completed with regard to attempting to provide a definition of the essence of EBMT compared with SMT and traditional RBMT. It highlights some important areas of EBMT still relatively underdeveloped, including evaluation, which we will come back to in Chapter 2. Going back to defining EBMT, Carl & Way (2003:xix) point out that EBMT is situated somewhere between RBMT and SMT, as many EBMT approaches integrate techniques from both paradigms.

The first mention of EBMT dates back to a paper presented by Makoto Nagao at a conference in 1981, published three years later (Nagao 1984). Around the same time as Nagao's paper, the DLT research group in Utrecht were carrying out similar work. Nagao's matching technique involves using a thesaurus to measure the semantic proximity of words; the DLT group talk about a 'Linguistic Knowledge Bank' of example phrases; likewise Sadler (1991) and Sadler & Vendelmans (1990) talk about a 'Bilingual Knowledge Bank', an approach which shares similarities with the other two, and lies within the EBMT paradigm.

However, in this context, we will discuss the work of Nagao. In his paper Nagao talks about 'machine translation by analogy principle', and identifies three main components of EBMT, namely (Somers 1999:116):

a) Matching source text fragments against a database of real examples
b) Identifying the corresponding translation fragments

c) Finally, recombining these to give the target text

We will provide an example to illustrate these three stages (taken from Armstrong et al. 2006b:4-5):

(Source text) (Target text)

(A) **I live** in Paris *with my wife* = (B) **Ich wohne** in Paris *mit meiner Frau*

This example is based on our approach to EBMT, which employs the Marker Hypothesis to segment phrases/sentences (input and saved examples). The EBMT system, MaTrEx, used in the current research is described in detail in Chapter 3. For this example we will simply say that (A) is the input string and (B) is the generated translation. (C) is the training corpus (previously saved examples). The data in (C) are then chunked using the Marker Hypothesis, with useful chunks and the corresponding target segments being extracted and stored for later use (D).

| Example (C) | |
|---|---|
| I live in Dublin | Ich wohne in Dublin |
| There's lots to do in Paris | Es gibt viel zu tun in Paris |
| I love going to the cinema with my wife | Ich gehe gern ins Kino mit meiner Frau |

| Example (D) | |
|---|---|
| I live | Ich wohne |
| in Dublin | in Dublin |
| There's lots | Es gibt viel |
| to do | zu tun |
| in Paris | in Paris |
| I love going | Ich gehe gern |
| to the cinema | ins Kino |
| with my wife | mit meiner Frau |

The translation process starts by searching the English side of the original training corpus in (C) to see if it contains the whole sentence (A). If it does not, we chunk the input sentence (A) into smaller constituents (E), again using the marker hypothesis, and search for these segments in the corpus of aligned chunks (D).

(E)

I live
in Paris
with my wife

Once these segments are located in the database (D), they are then recombined using a decoder to produce the final translation given in (B). With EBMT, the new example and its translation (A) + (B) are now stored in the training corpus in (C). That way, if this same sentence is encountered again, it can be retrieved in its entirety along with the corresponding target translation, without having to go through the stages of chunking and recombination again. At least one EBMT system, Traslán (Groves 2008), which translates between English and Irish, is used commercially.[14]

EBMT has often been linked with TM technology, and while it is true that some TM tools (e.g. Déjà Vu) integrate EBMT functionality, the major difference between the two methods is the fact a TM "is an interactive tool for the *human* translator, while EBMT is an essentially *automatic* translation technique or methodology" (Somers 1999:115). After locating a set of relevant example(s) in the database, the TM leaves the decision up to the translator to accept or reject the presented data. On the other hand, the EBMT system automatically selects the examples, and produces the output, with the human having no say in the process.

Way & Gough (2005b) comment that EBMT may be suited to areas relating to controlled language research. Sumita et al. (1990 cited in Groves 2007) also comments that EBMT systems still seem to suffer from problems of coverage, and are therefore suited to sublanguage domains. These comments support the use of an EBMT system in this study, as the language of subtitling could be regarded "as an entity in itself that can be shown to differ from the spontaneous, authentic discourse of everyday talk." (Taylor

[14] <http://www.traslan.ie/index.html> [Accessed 10 March 2009].

2006:1). As we outlined previously, subtitles tend to be shorter than regular sentences in a text document, as they usually try to mirror spoken dialogue, but in a more condensed form.

### 1.4.4 Repetition and Reusability

Repetition and reusability are two notions related to text reuse, and which are central to the approaches taken in both computer-aided translation (CAT) and contemporary machine translation (MT). However, they are also notions which are rarely problematised in the MT literature. It is generally assumed that high numbers of repetitions within a source text make it very suitable for translation using TM technology, and that translation of such documents will be easier and faster than those with lower levels of repetition. The same can be said for corpus-based approaches to machine translation. A major advantage often claimed in EBMT literature is that the overall quality of translation increases incrementally as the set of stored translations increases, which means that the chances of finding an exact match become greater as the corpus size increases (Somers 1999:92, Way 2003:443-444). It has been observed, however that there are two possible knock-on effects from this increase in corpus size: firstly, increased computational costs if the EBMT system stores examples as annotated linguistic structures, and secondly, the possible redundancy of identical or similar examples. For some EBMT systems the increase in SL repetitions is described as 'extra baggage', because the extra examples present the system with a kind of 'ambiguity' (Somers ibid). These observations call for further investigation of repeated SL segments and their corresponding TL translations. To date, repetition of source language segments and the reusability of target language segments within the scope of translation technology have not been researched to any great extent. Within the realm of TMs, the pre-analysis tool available with the TM software can give an indication of the number of repetitions which occur in the source text. This will tell the translator how many *exact matches* and how many *fuzzy matches* he/she is likely to encounter when translating a document. However, these repetition/match levels do not tell us anything about the possibility of reusing the same translation for each occurrence of the same source text segment, given the role context plays in all texts. The same can be said of CBMT. There is not usually a counterpart of the TM pre-analysis stage for Corpus-Based MT systems. The norm here is to use a bilingual corpus which is of the same or similar text type to the unseen input data, and to continually increase the size of the training corpus, with

the aim of improving the output. In addition, exact matches in the EBMT process are considered the exception rather than the rule, and translations are usually derived by breaking up the input sentence into smaller segments and recombining the translations of these smaller segments in the final stage (Turcato & Popowich 2003:73).

As already indicated, research in the areas of repetition and reusability in relation to translation technology is generally lacking. That said the following two studies highlight relevant points of interest relating to both areas, which are pursued further in the current research. The first of these studies was carried out by Whyman & Somers (1999), and describes a metric for evaluating TMs. The aim is to evaluate the translations retrieved from the database, and to indicate their "usefulness" to the translator. Whyman & Somers argue that the usefulness of a translation can be captured objectively by measuring "the effort required to convert the proposed match into the correct translation" and that "this effort can be quantified in terms of the number of *key-strokes* needed" (ibid:1274). They state that for this process of counting key-strokes to be accurate, one would need to count the number of key-strokes required to change the target text segment stored in memory into an 'ideal' target text translation. However, determining what an ideal translation is always introduces an element of subjectivity into the process, so in order to avoid this, Whyman & Somers decide that for "practical purposes" (ibid), an accurate measure can be derived by counting the number of key-strokes required to change the source language input to the source text stored in memory, i.e. counting the number of key-strokes to transfer from English into English.

Some points need to be made here in relation to Whyman & Somers' claims of objectivity. The metric they employ relies on *hits* and *matches*: hits being a specific term to indicate "a match deemed 'relevant'" and matches being a more general term to mean "any proposed retrieval from the database" (ibid:1272). When deciding on whether or not a segment is a hit, they take the highest ranked match for each segment, and make "a subjective evaluation of its usefulness for translation, based on previous translation experience" (ibid:1277), thereby introducing an element of subjectivity into their analysis.

A second subjective element in Whyman & Somers' methodology relates to the selection of TM. They stress the need to use a database that is 'appropriate' for their test data, remarking that the text (test data) they use in their experiment is appropriate to be

run against the database, and that it contains "a number of text segments for which matches in the database will be found" (ibid:1272). This indicates that they have some prior knowledge of their database contents, and the contents of the input texts. Within the current research, we investigate the value of having prior knowledge of the corpora the system is using, as it can give an indication of segments the system will deal with successfully, and those it might find problematic. What is very interesting to note with Whyman & Somers' study is that they immediately disregard any *exact matches* from their calculations, firstly commenting that "an exact match will be found which can then be simply pasted into the target document" (ibid:1266), and then later going on to say "exact matching of strings of characters is such a straightforward problem ... that this aspect of TM software is of no interest to us whatsoever" (ibid:1268). They thus exclude exact matches from their analysis. In his discussion on EBMT evaluation Somers (1999:147) notes that most evaluations of EBMT output exclude from the test set any exact matches with the database, as identifying these exact matches is seen as trivial. This omission is misguided in our opinion as the usefulness of the translation in exact matches also needs to be considered, in view of the particular context in which the given segment is used.

In the second of these studies, Reinke (2004) conducts a detailed evaluation of the retrieval performance of three TM tools. Like Whyman and Somers, Reinke notes that the 'relevance' of matches in a TM can be judged either on the basis of formal similarity between stored and new source language segments, or according to the extent to which the retrieved target language segment fits into the new, as yet emerging, target text. However, in contrast to Whyman and Somers, Reinke does not assume that one can automatically paste a target text segment into a new text if an exact match is found between the source text segments. He highlights two reasons for this: two sentences may be orthographically identical, but they may have either the same meaning or different meanings. If they have the same meaning, this is a source text repetition. Nevertheless, the ST segment might have different translations on different occasions for reasons of text cohesion and/or coherence, amongst others (ibid: 154ff, 237ff; see also Bowker 2005, López Ciruelos 2003, and Nedoma & Nedoma 2004). If the identical ST segments have different meanings, the segments are ambiguous. Reinke further believes that similar (as opposed to identical) sentences may still have the same meaning (meaning they are paraphrases of each other), and this could be helpful to the

translator, because when translators are faced with a new query sentence, they will normally want to know how a sentence with the same meaning was translated in the past.

Once again like Whyman and Somers, Reinke (2004:153ff) uses metrics familiar from Information Retrieval, namely precision, recall, and the associated F-score, which can be used to gauge the extent to which a system has retrieved all and only relevant matches from memory. He first ascertains repetition levels in new texts to be translated, and exact and fuzzy match levels between these new texts and existing translation memories. Although his evaluation procedure yields 'objective' numerical measures for each system's retrieval performance (in terms of recall and precision scores), Reinke (ibid:171) ultimately argues for a more qualitative approach to evaluation, and his analysis is actually characterised by detailed discussions of the actual contents of segments. Reinke is a firm believer that careful analysis of source texts and their translations is vital before we can assume that they will provide easily reusable translation equivalents (ibid:386).

The current research shares some similarities to both Whyman and Somers (ibid) and Reinke (ibid). Like the former, during the first stages of research and for the purposes of the prospective phase of the evaluation, we investigate a single solution for each segment, although the EBMT system we use can potentially produce several translations for a single input. Unlike Whyman and Somers, however, we focus on the target language, and we do not deem any of the non-preferred solutions offered by the EBMT system as noise, but rather we evaluate these qualitatively alongside the target language translations offered as first choice by the system. Like Reinke, we have a very deliberate separate source-text analysis phase, and we make predictions about how levels of internal repetition in our test data and 100% matches between our test data and training corpora might be expected to influence levels of reusability of translations in our corpus. Again like Reinke, we place heavy emphasis on the qualitative evaluation of translations proposed by our system, although we do not avoid quantitative analysis.

## 1.5 Previous Related Studies in Automated Subtitling

There have been previous studies on automatically generating subtitles and we discuss these in the following sections, dividing them up based on the type of technology used in the study. During our discussion of the systems, we outline the evaluation methodology used in the study.

### 1.5.1 Rule-Based Machine Translation

Nippon Hoso Kyokai (NHK, the Japan Broadcasting Corporation) and Catena-resource Institute joined forces to develop a Rule-Based Machine Translation system (STAR) that could be used to generate Japanese subtitles on news programmes from around the world. Since 1989 there have been two applications of STAR: firstly, to produce Japanese subtitles for English language news programmes (dissemination), and secondly, to produce rough Japanese subtitles for the newswire translation service (assimilation). The subtitles provided on the news programmes are usually presented in 5 minute slots, in a process that involves pre- and post-editing by Japanese translators. The newswire subtitles are real-time rough translations of incoming bulletins from an international wire service without any human intervention (Sumiyoshi et al. 1995:4). The STAR system uses a transfer-based approach to MT, comprising of four stages: morphological analysis, syntactic analysis, transfer and generation (Aizawa et al. 1990:308). Even though this system is commercially available, it has been specially adapted for use at NHK, i.e. making it more adaptable to translate news bulletins (Vasconcellos et al. 1991:123). During the syntactic analysis the system derives all the possible surface structures for an input sentence and the best candidates are then chosen by using a 'weight mechanism'. Weights are assigned to words and phrases based on nodes in an AND/OR graph and the corresponding rule. The smaller the weight, the better the candidate, as weights represent some kind of incomprehensibility or complexity of a word, phrase or sentence (Aizawa et al. ibid:310). Following a three month trial in 1991 of using this system to translate English-Japanese subtitles, Aizawa et al. (ibid) report that 64.5% of the news sentences were analysed correctly, and 78% of these were properly translated using the weight mechanism (meaning there was no need for post-editing). They note that the system had problems analysing the input sentences due to spelling errors and grammar mistakes present in the input sentences. However, these sentences are prepared by a bilingual translator from the original news. They also

point out that colloquial expressions are difficult to analyse. They note that the next stage of development is to focus on these weaknesses in the system. Since these results, developments with the system have included a method for selecting verb translations using machine learning theory that allows the system to learn decision trees from an English-Japanese bilingual corpus, and incorporating semantic categories, which reduces the proportion of unknown nouns to around 50%. This has meant an improvement of 5-10% for verb translation, and the error rate for verb translation is around 30% (NHK Annual Report 1994). In addition to system improvements, Sumiyoshi et al. (1995:4-7) outline improvements to the process of automatic subtitling, beginning with the source text input and ending with a preview of the subtitled news programme. A prototype subtitle generation system, or translation workbench, was developed to integrate all the tasks efficiently. This workbench allows the translator to input the original text into a word processor and to check the spelling, to machine translate the text and post-edit where required, to store the post-edited text in a file with the original English text, to insert subtitle cueing times, and to preview the programme with the post-edited subtitles, all of which is conducted before the final broadcast.

In 1994 NHK also began to conduct research into developing a Japanese-English MT system, as the demand for Japanese news programmes abroad increased. In addition to developing a RBMT system similar to the English-Japanese MT system, NHK began research into an Example-Based MT (EBMT) system. By 1996 NHK improved the speed of the RBMT systems through the use of decision trees, and the use of large bilingual corpora. They added a co-occurrence dictionary to the English-Japanese MT system to improve translation accuracy. Improvements were made to the Japanese-English MT system by further developing the transfer module and the bilingual Japanese-English dictionary. At this time research continued into developing an EBMT system with the creation of news data corpora, and an investigation into alignment techniques (NHK Annual Report 1995). Up until 2001, NHK describe the use of a Japanese-English RBMT system, and the improvements that have been made to the dictionary (increasing both function and content words) and the algorithm used to determine the English word order. They also introduce a machine-aided translation system, or translation workbench, which is used in conjunction with the MT system. This translation workbench includes a translation example browser, a term retrieval function, and a bilingual web retrieval function. This shares many similarities with the

well-known Translation Memory tools, such as SDL Trados' Translator's Workbench and STAR Transit.

In 2002 (NHK Annual Report 2002), NHK report on an investigation into what they call "pattern-based MT". They say that this system "works by matching an input Japanese sentence with sentences from a translation pattern database that stores patterns of Japanese to English translation" (NHK Annual Report 2002:27). This description seems to be another way of describing an Example-Based Machine Translation system. After testing the system, NHK report exact matches between 22% of the input data and the bilingual corpus, and claim that the target text translations were very accurate. They also obtained partial matches for 27% of the input data, and the translation accuracy for this MT output was estimated at 89%. There is no published information on the kind of evaluation NHK conduct to measure the quality of the MT output. However, in some respects, the viewers of the programmes which include subtitles can be seen as the evaluators of the output, and their feedback could be included in an evaluation methodology.

Since 2003, NHK's research into automatic translation has moved away from the original aim of providing Japanese subtitles on English language news programmes and now focuses primarily on machine-aided translation systems (in conjunction with an EBMT system) to translate text documents, and on speech recognition systems to generate intralingual subtitles for the deaf and hard-of-hearing (SDH) in collaboration with ATR Spoken Language Translation Research Laboratories (NHK Annual Report 2003, 2004).

Another RBMT system used to generate subtitles is presented by Popowich et al. (2000). They created a system (*ALTo*) to translate subtitles from English into Spanish on North American television. Their approach is based on Whitelock's (1994) Shake and Bake MT paradigm and also relies heavily on lexical resources. That said Popowich et al. consider their approach to be transfer-based, even though they have no structural transfer rules. The *ALTo* system has three distinct stages: a unification-based parser (including a proper name recogniser) analyses the input text; the transfer module maps source lexical signs onto target lexical signs, guided by transfer rules in the bilingual lexicon; these target lexical signs are provided as input to the generation stage, which

generates grammatical Spanish output (including correct word selection and word inflection).

Popowich et al. (ibid) argue that reading subtitles is very different to reading other translated texts, as the viewer only has a limited time to read and comprehend the subtitle. Therefore it is essential that translated output is grammatically correct. Based on this guideline, in some cases they omit certain "ungrammatical" elements from the source text segments without affecting the understandability of the entire sentence, to aid the RBMT system. Volk (2008:205) comments that this idea is "debatable" and it "opens the door for incomplete output". While grammatical output may be important, given that reading times are limited and grammatical output improves readability and comprehensibility (cf. James 2001), we would agree that omitting elements in the output is not the answer and would not support Popowich et al.'s argument that you can simply rely on other channels of communication to understand a foreign-language dialogue.

Although Popowich et al. comment that "the characteristics of the operational context influence the type of declarative evaluation that is appropriate", (ibid:334) the actual evaluation conducted is a declarative evaluation, as the evaluator rates groups of subtitles on two scales, without viewing the subtitles in an authentic AVT environment.[15] The declarative evaluation strategy follows those previously proposed in the literature (cf. Pierce et al. 1966, Nagao 1989, Arnold et al. 1994), evaluating the output based on two scales: grammaticality and fidelity. There are two points to note about this human evaluation: firstly, there is only one evaluator, and secondly, that evaluator evaluates groups of subtitles and not individual subtitles, since the meaning of one subtitle is often retrievable from the previous or next subtitle in the sequence.

Popowich et al.'s (ibid:337) interim results show "70% of the translations would be ranked as correct or acceptable, with 41% being correct", with an (experimentally) calculated estimation of 70% to 80% of correct/acceptable rankings following further development. The actual results are impressive and the expected results seem even more promising. However, given that the subtitles were evaluated by only one evaluator and the evaluation was text-based only, excluding the other media channels, it is difficult to

[15] Automatic evaluation metrics had not been developed at the time of this study. We will explain what "declarative" and "operational" evaluations are in section 2.1.1.

come to any conclusions on the success of this project. Popowich et al. (ibid) say they plan to conduct two more studies: firstly, an operational evaluation of MT output, in which comparisons would be made with human-generated subtitles; secondly, they want to compare the results of an operational evaluation of MT output with the results of a declarative evaluation of MT output. It is not clear whether these proposed evaluations were ever conducted or whether this system was employed commercially; however, the study raises interesting points in relation to a time-constrained translation domain such as subtitling, including "how the visual context can contribute to increase the acceptability of an incorrect translation" (ibid:336), which are not dealt with in other related studies.

The final example of using a Rule-Based MT system to translate subtitles is provided by Global Translation Systems (GTI), Inc. (Díaz Cintas & Remael 2007:20-21).[16] GTI use SYSTRAN[17] to translate English subtitles, in real time, into Spanish subtitles on selected television programmes. They conduct this process in conjunction with VITAC, a captioning and subtitling service provider. There is no published literature on the quality of the subtitles generated and therefore we are unable to comment on the success of this service to date. Díaz Cintas & Remael (ibid:21) comment on the "apparent lack of human agents and in the fact that this approach seems to be driven solely by economic forces and interests." Díaz Cintas & Remael (ibid) give examples of subtitles provided on the translatetv.com website and comment on the lexical and syntactic errors in these subtitles, adding that "if the examples shown on their website are meant to be the flagship of their trade, the situation becomes worrying." On the other hand, judging by the number of awards the company has received, the use of machine translation solutions in this area could be a very realistic option in the not too distant future.

### 1.5.2 Speech Recognition, TM and RBMT

Piperidis et al. (2005) describe the *IST/MUSA* (**Mu**ltilingual **S**ubtitling of Multimedi**a** Content) project, which took place from 2002-2004 and was funded by a consortium of companies. The project aimed at combining speech recognition, text analysis, TM tools and an MT system (SYSTRAN) to automatically generate subtitles. This is done by converting audio streams into text transcriptions (English to English); the transcriptions

[16] Global Translation Systems, Inc. <http://www.translatetv.com> [Accessed 10 March 2009].
[17] <http://www.systran.co.uk/> [Accessed 10 March 2009].

are condensed using text summarisation techniques to shorten sentences to the length suitable for subtitles (e.g., space and time constraints), and these new English subtitles are translated into French and into Greek using a combination of TM tools, an MT system and terminological resources. It is mentioned on the project homepage that the project involves human and automatic evaluation of the subtitles, but there is no detailed description of any evaluation process. Piperidis et al. (2004) mention that the acceptability of the automatically generated subtitles is currently rated at 45%-55%; however, there is no mention of any evaluation metrics. Given a lack of detailed information on Piperidis et al.'s evaluation methodology and the absence of sample output, it is difficult to comment on the validity of this assertion.

### 1.5.3 TM and Free Online RBMT

The following two studies use either TM tools or MT systems, or a combination of both to test the feasibility of automatically generating subtitles. Melero et al. (2006) present the *eTITLE* project, the goal of which is to highlight the potential application of combining existing TM tools and MT systems in the automatic translation of multilingual subtitles. The project works with the language pairs English-Spanish, Spanish-English, English-Czech, Catalan-Spanish, Spanish-Catalan, English-Catalan and Catalan-English.[18] The MT systems used in this study are free online rule-based systems. The study asks two main questions: is it better to use a TM tool in tandem with an online MT system to automatically translate texts, or to use only an online MT system? and, would translators save time if they translated subtitles firstly using *eTITLE* and then post-edited the results, compared with translating the subtitles from scratch? Melero et al. evaluated the output from a TM and MT combination and MT by itself using BLEU (Papineni et al. 2002) and NIST (Doddington 2002) scores. The usability evaluation showed that movie subtitlers would save time (approximately 17%) if they used the *eTITLE* facility and post-edited the output. Another benefit of using *eTITLE* is the correct lexical choices made by the system, thus reducing the amount of time spent by translators referencing dictionaries. That said, the evaluation compares texts of different lengths and it is not clear whether electronic dictionaries etc. were provided for the translators in the 'control' condition, making us question the real benefits of a 17% saving in time.

[18] When translating between Catalan and English (either direction), they use Spanish as a pivot language and therefore use results from the other language combinations.

O'Hagan (2003b) presents a preliminary study that asks whether language technology can respond to the new pressures being put on subtitlers, which include shorter time frames and increasing workloads. Three independent developments seemed to justify the need to examine the potential role of translation technology applications in subtitling: first, complaints by cinema-goers in Japan about the Japanese subtitles provided on the movie *The Lord of the Rings: The Fellowship of the Ring* (see also O'Hagan 2003a); second, the 'surprise language' experiments funded by the US Defence Advanced Research Projects Agency (DARPA); and, third the emergence of fan-subs, which refers to "the subtitles produced unofficially by Japanese animation (anime) fans for non-Japanese speaking viewers outside Japan" (op. cit.:2).

O'Hagan (2003b) investigates two subtitling scenarios translating from English to Japanese: in the first a Japanese subtitler, with a tight deadline, uses an off-the-shelf TM tool to see if it can help speed up the process of subtitling, and possibly improve quality; in the other an amateur translator uses a free online MT system to translate subtitles. This second scenario is based on the "fan-sub" model where the subtitler lacks formal training, but can draw on excellent genre knowledge combined with the MT output.

Scenario one presented many problems when it came to seeding the TM with English-Japanese translations. At the time of the study (and this is probably still the case), there was a lack of substantial English-Japanese parallel corpora in the field of AVT,[19] or indeed Japanese texts in digital form.[20] Therefore, O'Hagan built the TM using the English and Japanese subtitles from the first *Lord of the Rings* (LOTR) movie. The TM could then be used to begin subtitling the second movie in the trilogy. Subtitles are encoded on DVD as images, and thus Optical Character Recognition (OCR) software has to be used to extract them. The OCR software had particular difficulties reading the Japanese characters, and less significant difficulties reading English characters (mixing up i with l, for example). This slowed down the data capture process considerably. Once the TM was seeded with the subtitles from the first movie, the analysis data generated by Trados was not very promising, producing no exact matches, and fuzzy matches

[19] It may be the case that other languages are lacking; however, further investigations have been conducted since this study using English-German (Flanagan & Kenny 2007), Swedish-Danish & Norwegian-Danish (Volk & Harder 2007, Volk 2008), Dutch-English (Tiedemann 2007a, 2007b, 2008) and French-English (Lavecchia et al. 2007).

[20] The idea of scanning in the collection of *Lord of the Rings* books in English and the translations in Japanese was abandoned, due to problems encountered with the OCR technology and because the process was too time-consuming.

making up only 2% of the whole text. O'Hagan (ibid:11) notes that apart from short phrases and proper names, the "translation recycling idea did not work". She supports these comments with further observations regarding matching SL segments:

> Even where TM recognized a matching segment, the translation recalled from the translation memory was sometimes totally useless. This was due to the fact that some target language subtitles extracted from the memory were dynamic translations and were not applicable in a different context even though the source sentence may have been exactly the same.

The results from the second scenario showed that the number of Japanese characters in the human-translated subtitles was 55% that of the MT output, pointing to the condensing process conducted by human subtitlers. In comparison with the book translation, however, the human translation was 113% of the MT output, showing that humans produce longer sentences in the 'traditional' translation sense. The quality of the MT output was rated subjectively by an amateur subtitler, who would use this output to create good quality subtitles. Although O'Hagan (ibid) admits to the crudeness of the method employed, whereby the amateur subtitler judged whether or not each MT subtitle made sense on its own, without considering the context offered by subtitles before and after, or in a natural AV environment, the amateur subtitler deemed 80% of the LOTR subtitles intelligible. The same evaluation was conducted using *Harry Potter* subtitles, but the results were not so promising (50% rated intelligible), which seemed surprising given *Harry Potter* is aimed at a younger audience, but the system had problems dealing with the many French and Latin references (as did the human translator in some instances cf. Brøndsted & Dollerup 2004). The MT system fared worse when translating excerpts from the book, producing only 37% intelligible subtitles. The study was unable to test whether the MT output would be of any use to a non-translator genre-expert in producing good quality subtitles, as no suitable participant was available for the study.

This preliminary study highlighted numerous topics that needed further exploration, including translation recycling, corpus-analysis of TL segments, the issues of a non-translator genre-expert using MT output as a basis for creating subtitles (particularly in the area of fan-subs), and the many 'technical problems' encountered that could be

avoided to some extent during the MovRat project (outlined below) and even more so when conducting the current research.

### 1.5.4 Statistical Machine Translation (SMT)

The studies described in the previous sections all used a rule-based MT system. Following O'Hagan's (2003b) study, she suggested that different movie types and different MT system approaches should be used in future studies to compare the MT performance. Given that CBMT research is the most prevalent in the research community, it is no surprise that this approach would be applied to the area of subtitling with the aim of dealing with some of the problems associated with RBMT, including lexical coverage, modularity and robustness.

Volk & Harder (2007) and Volk (2008) present an SMT system for automatically translating television and movie subtitles from Swedish to Danish and Swedish to Norwegian, in order "to produce draft Danish translations to speed up the translators' work" (Volk & Harder 2007:499). They implement this system in a commercial setting as they conduct the research in conjunction with a large subtitling company in Stockholm. The motivation for developing an SMT system comes from wanting to "deliver a working system after a short development time and in order to best exploit the existing translations" (ibid:500). The SMT system is trained using GIZA++ (Och & Ney 2003) for the alignment, and Thot (Ortiz-Martínez et al. 2005) for phrase-based SMT, and uses Phramer, a phrase-based SMT decoder. Volk & Harder train their system on 4 million subtitles, taken from TV programmes, including soap operas, detective series, animation series, comedies, documentaries, and also from feature films, and their test set consists of 1,000 subtitles, randomly selected from the part of the corpus not used in training. In the first phase of the study they evaluate the SMT output against (independent) human translations of the same subtitles and obtain an average BLEU score of 57.3. This BLEU score is computed over all automatically translated subtitles. They then conduct a second evaluation which calculates the BLEU scores between the system output and the same output post-edited by six translators. This results in an average BLEU score of 65.8. They interpret this evaluation as showing that MT subtitle output plus post-editing yields good quality translations, and argue that calculating automatic scores using reference texts translated by humans does not provide a true picture of translation quality. Like Melero et al. (2006), Volk & Harder

(2007), and Volk (2008) find that post-editing of automated subtitles by a translator reduces the overall translation time, compared to a translator translating the subtitles from scratch. Volk (2008) reports that the customer they are working with is satisfied with the system output, and has recently started to employ the system in large-scale production.

### 1.5.5 Corpus Profiling

Corpus profiling is a method of investigating particular characteristics of a corpus being used to train a Corpus-Based MT system. To the knowledge of the researcher, this process is not usual practice in the MT research community. However, an example of corpus profiling is conducted by Volk (2008:209-210), when he investigates the vocabulary size of the corpus, noting lexical variance of Swedish and Danish words. He also investigates the repetitiveness of the subtitles, commenting that 28% of all Swedish subtitles in the training corpus were repeated. Of the repeated subtitles, half of them have exactly one Danish translation and the other half have at least two Danish translations. We also conducted corpus profiling in this study based on three characteristics of the corpora used to train the EBMT system: the number of source language repetitions they contain; the size of the corpus; and the homogeneity of the corpus. We calculated the percentage of SL repetitions for each corpus: 17% (1181 segments) of SL subtitles in Corpus A are repeated, 17% (1893 segments) in Corpus B are repeated, and 15% (6502 segments) in Corpus C are repeated.[21] Corpus A contains 6,997 aligned subtitles, Corpus B contains 11,342 aligned subtitles and Corpus C contains 42,331 aligned subtitles. This means that Corpus B is 62% larger and 38% less homogeneous than Corpus A; Corpus C is 505% larger and 83% less homogeneous than Corpus A. The aim of conducting corpus profiling is to become familiar with the content of the corpus in advance of using it to train the MT system. In this study collecting corpus profiles allows us to investigate possible relationships that exist between the profiles, on the one hand, and viewers' judgements of the intelligibility and acceptability of the subtitles produced by the EBMT system, on the other.

[21] The three corpora used in this study are called Corpus A, Corpus B and Corpus C. However, we refer to them as Corpus AM, Corpus BM and Corpus CM when discussing the machine-generated output, as these corpora do not include the *Harry Potter* test data when used as training corpora. This is explained later in section 4.1.

### 1.5.6 MovieTrans: Rapid, Memory-based Audiovisual Translations

O'Hagan's (2003b) study prompted further interest in applying CAT to AVT, resulting in the MovRat project, to which the present researcher made a substantial contribution.[22] The following description of the project builds on work reported in Armstrong et al. (2006c), and this work is used as a pilot study for the current research. This one-year Enterprise Ireland-funded project set out to test the feasibility of seeding an EBMT system with human-generated subtitles to automatically translate new movie subtitles from English to German and English to Japanese. The motivation for using EBMT over freely available RBMT systems was that a corpus-based approach allows the translator to build up resources to increase productivity, much like the rationale behind a TM. However, O'Hagan's work highlighted the limitations of TM technology for the current task.

This prompted the introduction of an EBMT system based on the Marker Hypothesis, meaning that segments are broken up into smaller segments based on a specific group of marker sets: determiners, quantifiers, prepositions, possessive pronouns, personal pronouns, conjunctions, wh-adverbs and punctuation. This results in "chunks" in SL and TL, which are then aligned. There are four main modules in the system: word alignment, chunking, chunk alignment and decoding. As in Volk & Harder (2007) and Volk (2008), the extraction of word-level alignments is conducted using GIZA++, a statistical word alignment tool. The decoder makes use of the Pharaoh phrase-based SMT decoder, essentially making the EBMT system a 'hybrid example-based SMT' system (Groves 2007, Armstrong 2007, Armstrong et al. 2006a, Stroppa & Way 2006, Stroppa et al. 2006; a more detailed description of the system is given in Chapter 3, section 3.2.2).

#### Corpus Creation

Data-driven approaches to MT rely on the availability of a sententially-aligned bilingual corpus on which to train the system to extract and store source-target sub-sentential alignments at a later stage (Armstrong et al. 2006c). O'Hagan (2003b) and Volk (2008) both mention the lack of availability of aligned subtitle corpora, and the shortage of relevant literature. Since the corpus-driven approach to automatically translating

[22] The present researcher's background is in both computational linguistics and translation studies, a combination that qualifies her to conduct the research presented in this thesis.

subtitles was novel, it was not clear whether a subtitle-specific homogeneous corpus would produce better results than a general language corpus made up of non-subtitle sentences. Therefore four corpora were created, two per language pair. The homogeneous corpus consisted of human-translated subtitles extracted from a collection of DVDs of English language movies which contained German or Japanese subtitles alongside English intralingual subtitles. In this pilot study movies were not selected based on genre, as the main priority was gathering data to train and test the EBMT system. As a result, Armstrong et al.'s subtitle corpus contained subtitles from thirty-six movies covering various genres including fantasy, action, romance and comedy. [23]

Due to time-constraints imposed on the project, the research team attempted to ensure the quality of the subtitles by taking them from major motion pictures, which tend to have high-quality subtitles. The extraction of the subtitles was also conducted by researchers competent in the given language combination, so that any errors spotted could be rectified at the corpus compilation stage.[24] Extraction of subtitles was conducted using the freeware SubRip.[25] This software uses OCR to convert the subtitles from image into text format, and once again problems like those outlined by O'Hagan (ibid) were encountered when dealing with Japanese characters. In addition, the availability of DVDs with Japanese subtitles in Ireland is limited and restrictions imposed on ordering DVDs from Japan due to region-code protection put on DVDs, thus prohibiting the sale outside the region, inevitably contributed to difficulties in gathering Japanese subtitle data. Consequently system development focused on using the English-German corpora to produce German subtitles.

After extracting the subtitles from the DVDs there was one text file of English subtitles and one (almost) aligned text file of German subtitles. The files were cleaned up by removing the time codes using Perl scripts, as time-codes would interfere with the training of the EBMT system. The alignment of the files was then verified manually by a researcher who had an excellent knowledge of both the source and target languages, in

[23] For a full list of movies used in Armstrong et al. (2006c) and in the current study, please refer to the Filmography.
[24] In the current study we are aware of issues of quality control related to corpus compilation and therefore conducted human-based quality checks (see section 3.4.1).
[25] <http://www.divx-digest.com/software/subrip.html> [Accessed 10 March 2009].

order to avoid any alignment errors.[26] Training an EBMT system on a corpus with numerous incorrectly aligned segments would have a negative effect on the MT output. Alignment is time-consuming, but by automatically numbering the lines using a text editor such as Word, the time spent on this process is reduced, as such numbering makes navigation through otherwise unsynchronized windows easier. Once the segments were correctly aligned, the numbers were removed from the segments and the file was saved as a plain text file.

There is an increasing interest in the area of creating subtitle corpora and automatically aligning the subtitles, despite the previous gap in the literature. Approaches include using alignment techniques based on time overlap and cognate recognition (which is superior to simply employing a statistical model of character length) (Tiedemann 2007a), a dictionary-based approach using automatic word alignment (Tiedemann 2008) and a technique known as Dynamic Time Warping that uses a bilingual dictionary to compute subtitle correspondences (Lavecchia et al. 2007). Volk & Harder (2007) and Volk (2008) align subtitles by using only subtitles with matching time-codes, suspecting that if the Swedish and Danish time codes differed by more than 0.6 seconds, the subtitles are not good equivalents. They say that by using this technique they "avoid complicated alignment techniques" used by the other researchers and that "most of the resulting subtitle pairs are high-quality translations of one another thanks to the controlled workflow in the commercial setting" (Volk ibid:209). The controlled workflow refers to the process whereby in-house Swedish subtitlers add their start and end time codes, and the Danish subtitlers normally use the same time codes. The Danish subtitler can of course change any time codes. In such cases the subtitle may subsequently be excluded from Volk's (ibid) corpus. It is assumed from Volk & Harder's description that no humans check the automatic alignments. That said, the alignment method seems reliable and practical but it is not realistically feasible outside of a commercial setting given the common lack of availability of bilingual subtitles.

The template-based approach (genesis files) mentioned in the Introduction shares similarities with Volk's (2008) approach mentioned above. This approach used for

[26] We originally aligned the subtitles following the removal of the time codes using automatic alignment techniques based on algorithms by Gale & Church (1991). However, on closer inspection of the aligned segments we noticed quite a few mismatched segments, some of which resulted from problems during the extraction of the subtitles from the DVDs.

DVD subtitling employs uniform in-time for all languages, which would aid alignment of such data.

The lack of available general language corpora for English-German also prompted the use of European Parliament proceedings for the heterogeneous corpus (Koehn 2005) in the MovRat project, as it provided a large bilingual corpus quickly and it is freely available.

There are of course some disadvantages to using the Europarl English-German corpus in the pilot study. The average sentence length of the homogeneous and heterogeneous corpora was calculated using WordSmith.[27] This showed the average length of subtitles to be a little less than nine words; in contrast the average sentence length in the Europarl corpus is 24 words, meaning the Europarl corpus had, on average, nearly three times as many tokens per sentence. There is also a significant difference in corpus size: the homogeneous corpus contained approximately 42,000 aligned subtitles, while the heterogeneous corpus contained in excess of 1 million sentence pairs. However, by using the two different corpora, we could test Denoual's (2005) claim that an EBMT system trained on heterogeneous data produces better results than one trained on homogeneous data.

The EBMT system was then trained using the two different corpora, output was generated and this output was scored using the BLEU automatic evaluation metric. BLEU provided a quick and easy way of ranking the translation output to investigate Denoual's (ibid) claims regarding homogeneous and heterogeneous training corpora. At the early stage in the research, it was clear that the BLEU scores did not reflect the quality of the output, but in all cases after training the system with the homogeneous corpus, the BLEU scores for the output were higher than when the system was trained using the heterogeneous corpus (see automatic evaluation below, Armstrong et al. 2006b). These results were further confirmed by Armstrong (2007). Consequently only the homogeneous corpus was used in subsequent evaluations in Armstrong et al.'s study.

[27] http://www.lexically.net/wordsmith/ [Accessed 10 March 2009].

### Evaluation Strategies and Results

Armstrong et al.'s study incorporated two approaches to evaluation: automatic and human-based. Within the MT community automatic evaluation is the norm. In contrast, within translation studies translation evaluation is usually conducted by human evaluators. The drawbacks of human evaluation, including the fact that it is time-consuming and costly, and based on subjective judgements, are acknowledged. However, as humans are ultimately the end-users of the output, human evaluation plays an important role in the development of natural language generation systems. Incorporating the two approaches resulted in a "balanced holistic evaluation of machine translation output" (Armstrong et al. 2006c:172). This holistic approach was divided into one automatic evaluation using the BLEU metric and three kinds of human evaluation.

*Automatic Evaluation*

The automatic evaluation was used to examine the type (homogeneous vs. heterogeneous) and size of corpus that would produce the best BLEU scores. Of the 42,000 aligned subtitles in the homogeneous corpus, 2,000 were selected randomly as the test data, and the remaining 40,000 were used for training data. A random sample of 40,000 sentence pairs was also taken from the Europarl corpus. Separate experiments were then conducted using subcorpora, beginning at 10,000 and incrementing by 10,000 on each occasion; for each subcorpus BLEU scores were calculated for the output. The motivation for this was to investigate the impact different datasets had on the output. BLEU scores were higher when the system was trained on homogeneous data (maximum BLEU score 10.8) than when it was trained on heterogeneous data (maximum BLEU score 5.8). Despite the relative increase of 86% between heterogeneous and homogeneous input, these BLEU scores remain very low. What is noteworthy from this automatic evaluation is the BLEU scores increase relative to the homogeneous corpus size, and this result is reversed for the heterogeneous corpus, suggesting that increasing the amount of non-specific data simply introduces more 'bad examples' (Armstrong et al. 2006b).

*Human Evaluation*

The human evaluation conducted can be split into formative (two examples) and comparative evaluation (one example). "Formative evaluation is designed to detect

areas requiring improvement while the system is still under development" (Armstrong et al. 2006c:174), while "comparative evaluation compares the performance between different MT systems in order to assess how the system under investigation fares against another MT system" (ibid). The first formative evaluation followed the classic human evaluation model of MT output, whereby evaluators rate individual sentences based on some scale. The EBMT system was trained on 30,000 subtitle pairs, and a test set of 2,000 subtitles was translated into German. 200 subtitles were randomly selected from the output, to avoid choosing the 'best' translations from the output. The 200 subtitles were split into four groups, and each of the four evaluators received a set of 50 subtitles. They were asked to evaluate each subtitle based on intelligibility and accuracy (cf. Pierce et al. 1966, Van Slype 1979), and provided with a four-point scale adopted from Wagner (1998). The scales ranged from 1 to 4, 1 being the best result. This evaluation is a very 'rough' approach, given that the subtitles were evaluated in isolation and they were not in sequence. The evaluation indicated the main areas of weakness of the system, including lexical errors, lack of capitalisation, problems in verb agreement, non-translated English words in the output, and problems with the chunking methods. The overall consensus from the evaluators was that the subtitles would need to be post-edited if they were to be acceptable for use on a commercial DVD.

The second formative evaluation introduced a relevant context to the process of evaluating subtitles. Six German native speakers individually viewed six 2-minute long movie clips which all had EBMT-generated German subtitles.[28] The evaluation sessions took place in a dedicated lab that simulated a home-cinema setup. The subjects could avail themselves of all channels of communication when viewing the clips. Clips 1-3 had an English language soundtrack, while clips 4-6 had a Japanese language soundtrack. The subjects' level of English was sufficient to watch and understand a movie; one of the subjects had only a very basic knowledge of Japanese, therefore Japanese was considered an unknown soundtrack language for all subjects. The group of subjects represented a homogeneous group, exhibiting the same language skills, balanced in gender and all attending a third-level institution.

[28] The number of subjects used in the MovRat pilot study is in line with recommendations in the literature (Carroll 1966:73, Van Slype 1979:181, Arnold et al. 1994:171, Dyson & Hannah 1987:166).

The subjects viewed all six clips before being asked questions during a retrospective interview. The interview[29] was administered by one of the project researchers and recorded on cassette tape. There were ten sections in the interview, each containing an average of four questions, to gather background information on the subjects, attitudes to subtitles, translation technology and machine translation, possible influence of source language knowledge on responses regarding acceptability[30] of subtitles, suitability of the subtitles, role of the image for comprehension, subtitle appearance and speed, and possible commercial developments. The interviews revealed that most of the subjects watched subtitled movies on DVDs three to four times a year, given that most movies shown in Germany-speaking countries are dubbed. Nonetheless, all subjects said that they preferred subtitled movies over dubbed ones, because being able to hear the original soundtrack gives a greater insight into cultural aspects of the movie. Two factors which could possibly influence the responses are familiarity with the technology and knowledge of the language of the soundtrack. None of the subjects were familiar with the technology used in the study, and there was a general consensus among the subjects that the EBMT subtitles without any post-editing could still benefit viewers who did not understand the soundtrack. In addition, they believed that if the subtitles were post-edited, they could be used in certain public situations, including film festivals and minority language scenarios, both of which may involve short release times and small budgets allotted for subtitling. They were, however, slightly hesitant to say whether they would accept these subtitles with post-editing on a commercial DVD. These responses show that a lack of familiarity with the technology did not negatively influence responses, and the feedback was constructive. Knowledge of the soundtrack may have influenced the subjects' responses, given that all subjects were slightly more critical of mistakes in the English source-language clips compared with the Japanese source-language clips. There could also be other explanations for this criticism: the present researcher observed that when the subjects were asked questions relating to all

[29] Using a questionnaire in a subtitling recipient evaluation builds on work by Gottlieb (1995:184), who uses a questionnaire to gather data on viewer reactions to deviations from subtitling standards. Unlike his study, Armstrong et al.'s (2006c) study and the present researcher administer the questionnaire to the subjects individually, ensuring all questions were answered. Gottlieb's study showed 22.1% of the questions on the questionnaire were unanswered or answered with "don't know", which is something we were able to avoid.

[30] Within the confines of the MovRat study the term acceptability is taken to mean sufficient quality, and a user would deem the subtitles acceptable if they were willing to use them while watching a movie. In Armstrong et al.'s (2006c) pilot study the concept was not investigated at a deeper level, for example with reference to FEMTI, which is the case for the current research (see section 2.2.2).

six clips, they nearly always commented on clips 1-3, recalling examples of errors, while even though they commented on errors in clips 4-6, they were unable to recall any concrete examples from these particular clips. Also there were examples in clips 4-6 where the subtitles were quite fast and there were numerous shot changes, often making it more difficult to take in sound, image and subtitles. However, the subjects frequently commented that the speed of the subtitles in clips 1-3 (which overall had fewer shot changes) was too fast and that the speed of the subtitles in clips 4-6 was satisfactory. These observations prompted the present researcher to investigate this point in the current study, by asking questions after each clip (giving the subject more opportunity to recall examples), and alternating the soundtrack between known and unknown language to examine the influence of the soundtrack language on the intelligibility and acceptability of the subtitles.

The final set of human evaluations was conducted through an online survey, using a Virtual Learning Environment (VLE), Moodle,[31] which is implemented campus-wide at the researchers' host university, DCU. The idea behind administering an online survey was to access a wider audience while keeping down costs and to test the technical ability of Moodle to host multimedia files and to allow access by potential participants. The type of evaluation was tested during the work of Armstrong et al. (2006c) to judge whether it could save time and costs to administer the evaluation sessions in this way, while at the same time generating useful data. However, using online surveys had its drawbacks, as it was more difficult to make sure that the subject completing the survey met the criteria that were in place to ensure the study was reliable and valid. This kind of MT human evaluation is not reported on in the literature, thereby making it important to test during a pilot study stage for the benefit of future studies in the area. DCU-registered students, native and non-native German speakers, with a good knowledge of German, could participate in the study. Despite there being a low response rate to the study, with the final number of participants totalling twelve, the evaluation process highlighted technical issues that need to be considered if a similar approach was to be developed on a larger scale. The subjects were once again asked to give background information, including information on how they normally watch subtitled media. Following this, subjects were asked to view six movie clips: three taken from *The Bourne Identityi* (2002) and three from *Harry Potter and the Prisoner of Azkaban*

[31] <http://moodle.dcu.ie> [Accessed 10 March 2009].

(2004).[32] All the clips had an English language soundtrack and one of three sets of automatically-generated subtitles: the first set contained raw EBMT subtitles; the second set contained subtitles translated using the free online MT tool, Babelfish;[33] and the third set contained post-edited EBMT subtitles. The Babelfish subtitles acted as a benchmark for the EBMT subtitles, given that it is probably the most well-known free online MT system, which in turn gives the subjects a good idea of the relative quality of the EBMT system. It was acknowledged, however, that the Babelfish system being an RBMT system, had not been trained on data similar to the test data, and therefore a direct comparison would be unjust. Post-editing was conducted by an English native speaker with knowledge of German, within a prescribed time frame of 20 minutes to post-edit 38 subtitles. The motive for this exercise was to test if non-native input could help to improve the quality of the subtitles, which could be a possible scenario where there are short release-times (e.g. film festivals).

For each of the three subtitle versions, subjects were asked to select the scenarios in which they thought the output would be acceptable. There were four different scenarios: purchased DVD, pirate DVD, in-flight movie and streaming video. For example, Table 1.2 outlines the results for a purchased DVD:

**Table 1.2**: Number of subjects who find the various subtitle versions acceptable for a purchased DVD

| Movie | Raw EBMT | Babelfish | Post-edited EBMT | None |
|---|---|---|---|---|
| Harry Potter | 1 | 1 | 9 | 3 |
| The Bourne Identity | 1 | 1 | 6 | 6 |

Armstrong et al. (2006c:177-178) give a detailed account of the results. Overall the subtitles on the *Harry Potter* clips were deemed to be acceptable more often than *The Bourne Identity* clips. For each of the scenarios, post-edited EBMT subtitles were deemed the most acceptable by subjects: purchased DVD (54%), pirate DVD (52%), in-flight movie (52%) and streaming video (53%). The post-edited subtitles received

[32] The reasons for selecting these two movies are that they both form part of a series of movies, and the researchers were considering investigating the benefit of EBMT when subtitling sequels, and that they are from different genres, which would be useful in investigating the suitability of EBMT for different genres.
[33] This system is powered by Systran <http://babelfish.altavista.com> [Accessed 10 March 2009].

positive comments from subjects, who commented that the subtitles "felt like German" (ibid:178). It is interesting to note the differing views of the subjects who watched raw EBMT clips with the researchers and the online survey group. The former were hesitant in saying whether post-edited EBMT subtitles are suitable for a purchased DVD, while the latter showed more confidence in the technology. This could of course be related to the presence of a researcher vs. the anonymity associated with online studies, and something that could be examined further. The raw EBMT subtitles were the least favoured, but the feedback from subjects regarding these subtitles was very helpful in relation to changes that could be made to the subtitles to improve acceptability. Babelfish subtitles were criticised for being too literal at times, and in some cases the register was incorrect. The human-based evaluation provided feedback that highlighted the shortcomings of the system from the end-users' viewpoint, something automatic metrics were unable to provide.

### Methodological Issues

The data gathered from this study provided many important insights into the acceptance of EBMT subtitles by end-users. However, given the relatively short timeframe for the project, some of the issues raised were unable to be examined further. Rather, the study established a basis for further research in the area. Reflecting on Armstrong et al.'s study, some comments are required in relation to the methodologies employed. The text-only evaluation approach followed the classic MT approach reported on in the literature. However, this approach is not necessarily suitable for the evaluation of subtitles, since the other channels of communication always present with subtitles were omitted, and the subtitles were lacking cohesion during the evaluation, given that they were not presented sequentially (cf. Popowich et al. 2000). The data gathered during the end-user evaluation sessions showed up a flaw in the consecutive mode of data collection. Showing the subjects six clips in a row and not alternating the soundtrack, and then administering the questionnaire, introduced problems of recall on behalf of the subjects, and this possibly biased the results, with all subjects focusing their criticism on clips 1-3, even though problems existed in clips 4-6. Another point relates to the acceptability of post-edited subtitles. The subjects who actually viewed raw EBMT subtitles and post-edited EBMT subtitles during the same evaluation session (online survey) were more convinced of the benefit of the latter compared with the group who were asked about post-edited EBMT subtitles, but who were not actually shown them.

All evaluations could have involved showing the post-edited versions, and gathering feedback on these during the retrospective questionnaire.

## 1.6 Concluding Remarks

This chapter has provided a comprehensive review of the literature in the areas the present study draws from, thereby establishing a sound basis for research into recycling texts from a translation technology point of view. By doing so, this chapter has provided a rationale for using EBMT to translate DVD subtitles from English into German. We have seen in the literature that subtitlers are coming under increasing pressure from the entertainment industry to provide high quality subtitles in ever-diminishing time frames, while at the same time pay rates for subtitlers are not increasing. Technology has been introduced into translation with relative success, most notably in the use of TMs by in-house and freelance translators and the use of MT systems by large organisations and research programmes. Translation technology has been shown to increase the speed of the translation process when used correctly and appropriately, allowing the translator to deal with the growing demand for translation. This chapter has shown the types of technology which the subtitler could choose from, namely Translation Memory and Machine Translation, and presented Corpus-Based MT as one solution for the translation of subtitles. Previous studies that use SMT and EBMT for the translation of subtitles were discussed.

The chapter has highlighted some gaps in the literature, most notably the lack of a large-scale human evaluation of automatically generated subtitles, which could be used as a benchmark for future research. Aside from the corpus profiling conducted by Volk & Harder (2007) and Volk (2008), there is also a lack of corpus analysis work on the corpora used to train MT systems, and investigating the importance of SL repetitions and the potential reusability of the TL translations in new contexts.

- Chapter 2
- Evaluation

# 2 Evaluation

Chapter 2 introduces the topic of machine translation evaluation (2.1), reviewing the kinds of evaluation already conducted in the literature (2.2), dividing these into human and automatic evaluation of translation quality. The chapter focuses in particular on machine translation evaluation within AVT, describes the model developed for this study, and the methods used to measure the pertinent quality characteristics of the MT output (2.3). Prior to the development and introduction of automatic metrics for MT evaluation, MT output was evaluated using human-based methods. The model presented in this study is also human-based, but it differs from the 'traditional' approach. We introduce a two-phase design: the first phase involves compiling a subtitle corpus, testing the quality of the corpus content, analysing the data in terms of repetitions (quantitatively) and reusability of translations (qualitatively); the second phase consists of end-user evaluation sessions of automatically-generated DVD subtitles. As a methodological starting point we use the Framework of Evaluation of Machine Translation (FEMTI) (Hovy et al. 2002a), an initiative of the International Standards in Language Engineering (ISLE), which provides a classification for the evaluation of MT and suggests evaluation methods that best suit a given context. This approach is combined with a recipient evaluation (Trujillo 1999), as FEMTI does not cover all of the aspects of the evaluation design developed for this study. Using these two frameworks we define the quality characteristics that are measured in order to examine the intelligibility and acceptability of EBMT-generated subtitles.

## 2.1 Machine Translation Evaluation

It would be nearly impossible to write a comprehensive review of the MT evaluation literature produced to date. The researcher would refer the reader to authors such as Hutchins & Somers (1992), Arnold et al. (1993), Falkedal (1994), King (1996, 1997), King et al. (1999), Hovy et al. (2002a, 2002b) and White (2000a, 2000b, 2003) for a comprehensive overview of many studies conducted. However, this section will mention some significant studies on MT evaluation, outline the importance of evaluation within the domain of MT research, and detail the scope of MT evaluation.

MT evaluation has been systematically conducted since around the 1950s, and as a result has generated an abundance of literature. It was Wilks (Carbonell & Wilks 1991)

who remarked that more has been written about MT evaluation than about MT itself. Even so, there are two observations that can be made in relation to this: firstly, King et al. (2003) point out that much of this literature is difficult to obtain, citing the example of the Van Slype (1979) report, a copy of which was extremely difficult to access prior to 2003; secondly, there is still no simple answer to the question of what the best method of evaluation really is. One problem as outlined by King et al. (ibid) is that often when researchers design an evaluation methodology, they believe theirs is of special importance, and design much of the new evaluation strategy from scratch, thus wasting reusable resources and time which could be spent on developing an already existing methodology further. Such concerns led to the setting up of initiatives such as EAGLES, ISLE and FEMTI and these are discussed in detail later on in this chapter.

King (1996) notes that an important task of conducting any evaluation is to take into account what is to be evaluated, and many researchers find this task difficult at times. By knowing what is to be evaluated, the evaluation methodology can be shaped to focus on a particular aspect of evaluation, instead of evaluating unnecessary aspects of an MT system. Since the late eighties and early nineties, there has been a push towards developing MT evaluation as a standalone strand in Natural Language Processing (NLP). However, problems have arisen with this; it is quite difficult to develop an objective methodology for MT evaluation which can be used to highlight improvement in particular approaches and particular systems, and show the advantages of one approach over another, given the subjective nature of human-based machine translation evaluation. This has led a drive towards standardisation in relation to MT evaluation methodologies and to a growth in automatic approaches.

As White (2003) points out, the aim of evaluation is to measure some attribute of something against a standard for that attribute. However in the case of translation, there is not just *one* correct translation, as if this were the case the problems associated with MT and MT evaluation would be long solved. Human translation of human languages cannot provide a *correct* translation, the necessary component for evaluation, a fact that stems from "the rich variability of language and remarkable creativity that goes into the act of translating" (White 2003:213). This could lead us to believe that the prospect of MT evaluation is ultimately impossible. However, this is fortunately not the case and the answer lies in being able to control the factors that can be controlled, and to be

consistent with other factors that cannot be controlled completely. This point is illustrated in the Methodology chapter (see section 3.1.2) when we discuss our research design.

### 2.1.1 General Approaches to Machine Translation Evaluation

There are a number of typologies of MT evaluation approaches currently in use. White (2003), whose work is based on that of Arnold et al. (1993), and augmented by the models of Van Slype (1979) and Vasconcellos (1992), organises MT evaluation into six main types:

- Declarative evaluation: measures the ability of an MT system to handle texts representative of an actual end-user (White 2003:227)
- Operational evaluation: addresses the question of whether an MT system will actually serve its purpose in the context of its operational use or if focusing on cost in particular, determines the cost-effectiveness of an MT system in the context of a particular operational environment (ibid:231)
- Feasibility evaluation: provides measures of interest to researchers and the sponsors of research of whether the system has any actual potential for success after further research and implementation (ibid:222)
- Internal evaluation: occurs on a continual or periodic basis in the course of research and or development to test whether the components of a system work as they are intended (ibid:224)
- Usability evaluation: measures the ability of a system to be useful to people whose expertise lies outside MT *per se* (ibid:230)
- Comparison evaluation: measures some attribute of a system against the same attribute of other systems (ibid:235)

The EAGLES Evaluation Working Group (EWG) distinguishes three types of evaluation (King 1997:252):

- Adequacy evaluation: assesses whether a system fulfils adequately a set of specific needs

- Progress evaluation: assesses whether a system has made progress towards some desired goal state of the same system. This type of evaluation is also referred to as internal evaluation, as defined above
- Diagnostic evaluation: finds out where a system fails and why

According to the EWG, the latter two can be seen as special cases of adequacy evaluation.

Throughout the literature there are some distinctions between approaches to MT based on how evaluations are conducted. Hovy et al. (2002a) and White (2003) discuss context-based evaluation as a type that relates the evaluation methodology to the purpose and context of the system. Another type of evaluation is task-based evaluation and this relates the evaluation methodology to a particular task set out by the evaluators and tests to see if the user can complete this task adequately by using the MT output (White & Taylor 1998, Doyon et al. 1999, White 2000b). A third type of approach to MT evaluation is reference-based evaluation, which involves the use of automatic metrics, for example BLEU (Papineni et al. 2002, and below) to estimate the quality of automatically translated sentences based on their similarity to human-translated versions of the source text.

### 2.1.2 Scope of MT Evaluation

Evaluation is related to an interested party wanting to know more about a particular component or components or aspect of the system. The interested party can be any number of people including developers of the system, researchers using the system or helping to design it, users of the system at various stages, users of the output and purchasers. Given that there are many different interested parties, the scope of MT evaluation is wide and varied. Some might only want to evaluate the speed of an MT system or compare different systems by cost and nothing else, and after narrowing down their choice based on cost, they might want to widen the scope and carry out a subsequent evaluation this time taking into account the cost, speed and coverage of the remaining systems (Arnold et al. 1993:4, White 2003:222). As mentioned before, the most important thing when designing and carrying out an evaluation is to know what it is you are evaluating or what you want to evaluate. Later, when we discuss the current human-based approach, we define the scope of our evaluation.

### 2.1.3 Large-scale MT Evaluation Studies

Throughout the history of MT evaluation there have been a number of prominent studies which have influenced the development of MT evaluation methodologies. Some of these studies can be considered as experiments, while some of them are frameworks which are established in order to evaluate MT in a systematic way. In the next few sections we highlight studies which have had an impact on the way evaluation has been conducted in the past and on current approaches to MT evaluation. They are divided up into human evaluations, automatic metrics and general frameworks for evaluation.

#### 2.1.3.1 Human Evaluation of Translation Quality

##### The ALPAC report: Carroll 1966

In 1966 the Automatic Language Processing Advisory Committee (ALPAC) (Pierce et al. 1966) published a report making nine recommendations for future research projects in translation. Of the nine recommendations outlined in the report three related specifically to further work within MT evaluation: develop new evaluation methods, evaluate quality and cost of translation, and evaluate the speed and cost of machine-aided translation. Even though this report was fairly brief, it proved to be detrimental to MT research, with some people even today still referring to it when they talk about MT and its apparent uselessness. For many, the report implied that MT had been a failure, or at best, was very unlikely to be a useful technology (Hutchins 2003a:133). The ALPAC report was a turning point for MT research in the US, with funding being withdrawn from many of the projects which were underway at the time. Hutchins's (ibid) paper on the report gives a very succinct account of the important points made in the report, and relates these to the development of MT research despite the report's findings, in particular research throughout Europe and Japan. He makes two important points in his review: firstly in hindsight it can be agreed that ALPAC had every right to believe MT research was going nowhere – the quality was unquestionably poor, and given the amount of time and money spent on developing this poor quality, it was not unreasonable that the report would question the justification for MT research. Secondly, however, one area which can be faulted is the way the report focuses exclusively on the translation needs of US scientists and of US agencies from Russian into English, without recognising the broader needs of potential users of MT systems. This also means that the report relates to MT in a specific context and is not concerned

in any way with other potential uses or users of MT systems, or with any other languages.

One evaluation study of particular interest was conducted by John B. Carroll (Pierce et al. ibid:67-75), whose goal was to establish standard procedures for translation quality. The aim was to make these standards applicable to human-produced translations and machine-translated texts. Carroll claimed (ibid:67) that previous attempts to evaluate translations, whether human or machine translations, were "too laborious" and "too subject to arbitrariness in standards, or too lacking in reliability and/or validity." Carroll therefore devised an experiment to evaluate sentences translated from Russian to English. There were six different sets of 36 sentences selected at random from four different passages (three sets translated by humans and three sets translated by a machine), and placed in random order within the test sets. Each set was given to three monolingual and to three bilingual speakers.[34] This meant that each set contained a different translation of a given sentence, and therefore no rater evaluated more than one translation of a given sentence. These measures reduced any rater bias. Each rater judged the translated sentences based on three criteria:

1. intelligibility (without reference to the original)
2. fidelity in relation to the original sentence
3. reading/rating times for each sentence

Intelligibility was measured on a 9-point scale, and for each point on the scale a description of the quality of the translation was provided. This ranges from 9 being "perfectly clear and intelligible" to 1 being "hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence." (Pierce et al. ibid:69). There is no actual definition of intelligibility provided in Carroll's study. From what the researcher can tell, the definition must be inferred from the descriptions on the scale, meaning intelligibility relates to how well the evaluators understand the translation. Fidelity, however, was measured indirectly through informativeness on a 10-point scale. In the study, in order to evaluate the fidelity of a translation, the evaluator was asked to gauge from the scale how much more information could be gathered from reading the original or human (reference) translation than from simply

[34] The monolingual subjects were native English speakers, and the bilingual subjects were native English speakers with an excellent knowledge of Russian.

reading the translation. If the original sentence was highly informative relative to the machine translated sentence, this meant that the translation was lacking in fidelity.

The results from the study showed that the ratings for intelligibility and fidelity are very highly correlated, when averaged over sentences and raters. The mean reading times also show a linear relation to the mean ratings, supporting Carroll's scale design. Two other findings also emerged from the study: to obtain reliable mean ratings, a fairly large sample of sentences need to be rated (no size suggested); and to avoid too much inter-rater variation, at least three or four raters should be used in this kind of study.

The Likert-type scale is one of the most popular types of interval measurement scales. It was developed by psychologist Rensis Likert (1932) to identify "the extent of a person's feelings or attitudes toward another person, event, or phenomenon" (Frey et al. ibid:103). Human evaluations of MT output following the ALPAC report all incorporated some form of Likert scale, sometimes using nominal or ordinal scales, but using them for the same purpose of measuring quality characteristics of translated texts.

### Van Slype Report

In 1979 Georges Van Slype compiled a comprehensive critical review of MT evaluation methods on behalf of Bureau Marcel van Dijk for the Commission of the European Communities, who had set up a programme aimed at "lowering the barriers between the languages of the Community" (Van Slype 1979:11). The purposes of this study were: to document the kinds of methodologies being employed at this time in MT evaluation; to make some recommendations to the Commission, amongst other things, on the methodology it should use when evaluating its machine translation systems; and to conduct research which would help in the long term with the efficiency of these evaluations. The report distinguished between two levels of evaluation: *macroevaluation* (or total evaluation) determines the acceptability of a system, compares the quality of two systems or two versions of the same system, and assesses the usability of a system; while *microevaluation* (or detailed evaluation) determines the improvability of a system. Within these two levels, there are four and five groups respectively. The Van Slype report outlines the many different kinds of scales previously used to measure criteria within the groups, and in the case of intelligibility and fidelity, the interval scales range from 2-3 point scales up to 25-point scales. In contrast to Carroll, for example, Van Slype uses a 4-point interval scale for both

intelligibility and fidelity. When measuring intelligibility Van Slype uses sequences of texts, (compared to the random selection of sentences preferred by Carroll), with between 5,000-10,000 words extracted from 20 to 40 documents. He states that the evaluator of the texts should be able to understand the context from simply reading a sample of the text. (Carroll, however, did not report any problems related to lack of context in his experiment).

Van Slype (ibid:54) is critical of Carroll's technique of measuring intelligibility and fidelity as the principal criteria of quality, as he says the two criteria are theoretically independent of each other. He also rates the effectiveness of fidelity assessment as poor as a method of evaluating MT, as the evaluator needs specialized knowledge of a subject-specific text, and even then evaluators' judgements will vary depending on the importance they attach to each sentence. Van Slype (1979) provides two definitions of intelligibility in relation to translation output, one by Halliday and the other from his own work. Halliday (cited in Van Slype 1979:62) defines intelligibility as the "ease with which a translation can be understood", but fails to provide a method by which this attribute can be measured, while Van Slype (ibid:62) defines intelligibility as a "subjective evaluation of the degree of comprehensibility and clarity of the translation" and, like Carroll, proposes a subjective rating scale to measure it. In this study we understand intelligibility as including both 'comprehensibility' (the ease with which a translation can be understood) and 'readability' (the ease with which a translation can be understood in a prescribed amount of time). We discuss these quality characteristics in more detail later in the chapter.

In relation to the two levels of evaluation outlined in the Van Slype report, we situate the current research in *macroevaluation*. The criteria proposed for this level of evaluation can be classified into four groups: *cognitive*, *economic*, *linguistic*, and *operational*. The level relevant to our research is the cognitive level, which involves the "effective communication of information and knowledge" (Van Slype ibid:57). There are five criteria associated with cognitive level evaluation, namely *intelligibility*, *fidelity*, *coherence*, *usefulness* and *acceptability*.

We defined intelligibility and fidelity in the sections above. To define *coherence* Van Slype (ibid:78) draws on the work of Wilks (1978). According to Van Slype (ibid), Wilks is quite vague in his definition of coherence. Wilks states that the quality of a

translation can be assessed by its coherence without having to study its correctness as compared to the source text, and that coherence can be inferred from a large sample of translations by a monolingual target language speaker. He believes that the probability that a large sample of translations is both coherent and very wrong is very weak. One thing to note here, however, as Van Slype rightly points out, is that Wilks's definition does not include any method for evaluating the coherence of a translation. Wilks believes a monolingual evaluator can rate the coherence of the target text, but then says that the coherence of a target text must be relative to the coherence of the source text, a different claim which could presumably only be substantiated by a bilingual evaluator.

According to Pankowicz (1978, cited in Van Slype ibid:33) the *usefulness* of a translation is based on quality, speed and cost, and determining the optimal balance between these three attributes depends on the context of use of the MT system. This definition of usefulness is supported by the work of Church & Hovy (1993). The Van Slype report provides various subjective methods (however, none are suggested by Pankowicz) to measure the usefulness of the target text using *n*-point scales.

Van Slype defines *acceptability* as "a subjective assessment of the extent to which a translation is acceptable to its final user" (ibid:92). Van Slype maintains that acceptability can be effectively measured only by a survey of final users and this is illustrated in his suggested subjective evaluation, the second of two methods for evaluating acceptability in the report:

1. Measurement of acceptability by analysis of user motivation, and
2. Measurement of acceptability by direct questioning of users.

The first method, used by Dostert (1973, cited in Van Slype ibid:93), means that users of MT are asked several questions dealing with their motivation, for example:

- Why do you use MT?
- How much MT do you request per year?
- What is the reason for which you use MT (cost, speed, confidence, exactitude)?
- Do you recommend MT to your colleagues?

The second method, used by Van Slype, has two steps. Firstly it entails submitting a sample of MT with the original texts and the corresponding human translations to a

sample of potential users, followed by a question session among the users of the MT output, including such questions as (ibid):

- Do you consider the translation of these documents to be acceptable, knowing that it comes from a computer and that it can be obtained within a very short time, of the order of half a day?
    - in all cases
    - in certain circumstances (to be specified)
    - never
    - for myself
    - or certain of my colleagues
- Would you be interested in having access to a system of machine translation providing texts of the quality of those shown to you?

Van Slype (ibid:112) also mentions, in general, advantages and disadvantages of end-users measuring the acceptability of MT. Advantages are that the judgement is made by the one for whom the translation is done, the evaluation criterion is simple – a text is either acceptable or it is not, and the measurement relates to the actual purpose of the operation (acceptance or not of the translated text by the user) and not to an intermediate or partial aspect (intelligibility, fidelity, etc); although, that said, the users' judgements do include these elements. The possible disadvantage of the method is that it deals with users with varied aims and perhaps a wide range of different document types. The report suggests that in order to obtain conclusive results, it is necessary to use a fairly large sample[35] of users and of documents, and the method then becomes very expensive. This means that within the framework of a macroevaluation and on a limited budget (given that a macroevaluation means a total evaluation and takes in three large areas of consideration when evaluating a system, see above), the method can only cover a small sample of the population and will thus be of indicative value only. However, if this method is used on a larger scale, it goes beyond the limits of a macroevaluation, and the operation becomes one of research (ibid:12).

Examining the two methods of evaluating acceptability outlined above, we can see that our methodology could not be based on Dostert's idea. The viewers of subtitles do not necessarily request MT output (not in our research in any case), and therefore this

[35] It is not indicated in Van Slype (1979) the number of users that would make up a fairly large sample.

methodology is not relevant or applicable in this context. On the other hand we can adapt Van Slype's method for measuring acceptability by incorporating user surveys/questionnaires. Because we conduct our evaluation on a medium scale, given the number of participants we gathered for our end-user sessions, the results are more than just of indicative value. We would disagree that acceptability can be judged on the basis of MT output being acceptable or not acceptable, and see acceptability as a more nuanced concept with a number of factors contributing to it.

Van Slype's method to evaluate the acceptability of MT output shares many similarities with the work of Bowker & Ehgoetz (2007). Drawing on Loffler-Laurian's (1996:69, cited in Bowker & Ehgoetz ibid:212) contention that "the recipients of MT output are in the best position to judge whether or not this output satisfies their requirements", Bowker & Ehgoetz set out to investigate whether recipients of certain types of machine translated texts would accept lower quality translations in order to save time and costs, or whether they would prefer to receive higher quality human translations, even if it means higher costs and slower turnaround times. They refer to an approach taken by Chesterman & Wagner (2002:81), who suggest viewing translation as "a service, intangible but wholly dependent on customer satisfaction." Therefore, to measure user acceptability, we need to measure customer satisfaction. Trujillo (1999:255) refers to this approach as recipient evaluation. He (ibid) points out that recipients of translations are usually monolinguals in the target language and says that their main concerns are with cost, speed and linguistic quality of the translation. Of course this list of concerns will not apply to every evaluation, as each evaluation is conducted for different reasons.

Other studies that investigate the 'acceptability' of MT output include Roturier (2006) and Coughlin (2003). Roturier (ibid) uses a two-phase approach to determine whether controlled English rules can have a significant impact on the usefulness, comprehensibility, and acceptability of MT technical documents from a Web user's perspective. He conducts an online experiment using a customer satisfaction questionnaire. To our knowledge this is the only study to date that situates an MT evaluation in a real-world setting and evaluates the three quality characteristics based on end-user responses. Roturier's use of acceptability is based on the fourth standard of textuality defined by De Beaugrande & Dressler (1981:7). The latter's use of acceptability refers to the "text receiver's attitude that the set of occurrences should

constitute a cohesive and coherent text having some use or relevance for the receiver." The current study shares with Roturier a concern with real-user evaluation. However, in contrast to Roturier, we simulate our real-world setting for a number of reasons which are discussed in detail in the Methodology chapter (see section 3.5).

Coughlin (ibid) focuses on the correlation between automated and human assessment of MT quality. She asked evaluators to use a 4-point scale to measure the 'acceptability' of MT output. Coughlin's approach aims at evaluating two quality characteristics simultaneously, namely intelligibility and accuracy. However, her approach can be criticised on a number of counts: firstly, the points in the acceptability scale used in her study are labelled by terms such as 'ideal', 'perfect translation', 'comprehensible' and 'accurate'. All of these terms are somewhat subjective, and the concepts behind some of them, for example 'perfect translation' simply do not exist (White 2003). Secondly, there is a difficulty with the use of the term 'accurate'. What might be very accurate for one evaluator might be very erroneous for another. Finally, Coughlin attempts to measure the acceptability of sets of translations produced by MT systems, without taking the users of the MT output into account. Coughlin (ibid:64) further notes in the acceptability scale that a translation can be considered possibly comprehensible if the evaluator is "given enough context and/or time to work it out…and some of the information is transferred accurately" from the source text (ST) to the target text (TT). Users (as opposed to evaluators) would not have access to the ST, and therefore would be unable to work out whether accurate transfer had occurred, no matter how much context or time they were given. This is particularly pertinent to the case of subtitling. Viewers of subtitles have only a fixed pre-determined duration to comprehend the subtitle. If users of the subtitles understand the original soundtrack, they could understand the subtitles by using an extra semiotic channel. However, the primary target audience of interlingual subtitles are users who do not understand the original soundtrack (Gambier 2003), and therefore use the subtitles to understand the movie.

The studies outlined here allow us to define acceptability as the satisfaction of the recipients' requirements, and they also point to the suitability of integrating a recipient evaluation into our evaluation model. In addition, they point up the increasing awareness in AVT of the need for reception studies. We note also here that Van Slype, one of the main proponents of recipient evaluations, maintains that in evaluations

focusing on final users of raw MT with acceptability as the main aim of the evaluation, reading time should be taken into account. Popowich et al. (2000) also make this point in relation to evaluating subtitles. This is incorporated in the current study through the use of the readability measurement. We consider reading time during the discussion of retrospective phase results in Chapter 5.

### DARPA/ARPA[36] MT Evaluation Methodologies

White et al. (1994) summarise evaluations conducted by The Defence Advanced Research Projects Agency (DARPA), a major US funding agency, which ran from 1991-1994. The aim of the DARPA evaluations was to examine the progress of their own MT systems. For all evaluations, 30 monolingual native English speakers made judgements about intelligibility and fidelity on 30 sets of texts translated from French, Spanish and Japanese into English, following an established standard of assessment that included:

- **Adequacy** (fidelity measure), required the subjects to indicate on a 5-point interval scale whether the MT output contained the same information as the professional human translation
- **Fluency** (intelligibility measure), required subjects to determine if the translation could be considered "good English", indicating this on a 5-point interval scale and without access to a reference (human) translation
- **Informativeness** (fidelity measure) was measured using a comprehension-like test, to determine if the MT output contained enough information for the subjects to answer multiple-choice questions

Measures were put in place to control for maturation and testing effects.[37] These included having only one article covering a particular news topic and ensuring subjects judged intelligibility and fidelity on different translation passages, so that, for example, the adequacy result would not influence the fluency result for that passage. As outlined above, fluency measures were conducted without access to a human translation, reducing any bias.

[36] This agency was called the Defence Advanced Research Projects Agency until 1993, and then changed to the Advanced Research Projects Agency. It once again reverted back to the Defence Advanced Research Projects Agency in 1996.

[37] These effects, along with other types of effects that can bias a study, are discussed in detail in section 3.3.4.1.

From the human-based studies outlined above we can see that fluency and adequacy (also referred to as intelligibility and fidelity; cf. Pierce et al. 1966, and Van Slype 1979, who uses 'fluency' and 'intelligibility', and 'adequacy' and 'fidelity' interchangeably), are the most common attributes of MT output that are measured in order to establish the 'quality' of MT translations. Before we discuss automatic metrics used for evaluation, we will discuss a different kind of approach used by JEIDA and the frameworks developed with the aim of standardising evaluation methodologies. In our discussion on automatic metrics we will return to the most widely used quality characteristics of fluency and adequacy.

### The JEIDA Survey

The following section looks at a slightly different approach to MT evaluation. Instead of using humans to evaluate output on rating scales, it discusses possible evaluation methodologies that can be used depending on the user of the MT system and the purpose of the system. In the 1997 Survey of the State of the Art in Human Language Technology, King (1997) includes an article on human factors and user acceptability. She comments on how little attention had been paid up until then in the published literature to users within evaluation in general, and goes on to mention essentially three classes of user: researchers or manufacturers concerned with system development, funding agencies – especially, in this context, (D)ARPA – and potential purchasers of commercially available systems. She is, however, quick to note one exception, namely from the area of MT. The Japan Electronic Industry Development Association (JEIDA) published a report in 1992 on evaluation criteria for MT systems. The evaluation methodologies developed by JEIDA focus on the fact that the different users of MT want to see different strengths of an MT system. JEIDA devised a comprehensive set of questionnaire materials that covered several views of the needs of the various stakeholders in MT (users, developers, production managers, etc). The questionnaires focused on three specific types of evaluation: user evaluation of economic factors (*Should MT be introduced into an environment and if so which type of system would be the most economical?*), technical evaluation by users (*If MT is introduced into the environment, which system would best fit the needs of the environment?*), and technical evaluation by developers (*Does the MT system meet the original objectives set out for coverage, accuracy, ease of use, etc?*). To establish the criteria relevant to the user evaluation of economic factors, the users answer two questionnaires: one to establish

the current translation situation, and the other to establish the users' translation needs (Nomura & Ishara 1992:11). The questionnaires covered areas such as current translation situation, organisation and turnaround time required. These questions were then associated with fourteen parameters, presented in Table 2.1 below, characterising MT systems that can be used to evaluate the users' answers objectively (objective ratings) (White 2003:233):

**Table 2.1:** Fourteen parameters which characterise MT systems

| | |
|---|---|
| A1 | Present translation needs |
| A2 | Type of document |
| A3 | Quality of translation |
| A4 | Language pair |
| A5 | Field of application |
| A6 | Time |
| A7 | Automation |
| A8 | Organisation |
| A9 | Cost |
| A10 | Text data extraction |
| A11 | Re-insertion of text data |
| A12 | Installation conditions |
| A13 | Pre-editing |
| A14 | Post-editing |

(Source = White 2003:233)

A value for each parameter is derived from the users' answers (numerical ratings). These values can then be represented visually in the form of a radar chart (visual judgement). MT systems are classified according to their type (there are seven different types), and each type has properties that correspond to the same fourteen parameters outlined above. This means that each type of MT system can also be represented visually in the form of a radar chart. This results in the easy task of matching a user's situation with a system type by comparing the configuration of the radar charts.

The second type of evaluation proposed by JEIDA enables potential users to evaluate the technical capabilities of a system (Hutchins 1993:25). Once again two questionnaires are used: the user fills out one questionnaire outlining the particular factors that are important in relation to the intended end-use of the system (including

quality of translation, introduction costs, and dictionaries). The second questionnaire is filled out by an MT system provider. This questionnaire is a self-assessment and preliminary evaluation of a system's output (White 2003:234). JEIDA provide formulae for the composition of questionnaire scores that are sensitive to the different intended end-uses. The result is a radar chart that represents two things: the performance of a particular system and the particular user's satisfaction with the system (ibid). The closer the peaks and valleys are on the chart, the more suitable the system is for the end-user.

The third type of evaluation proposed by JEIDA is not aimed at the user, but rather at researchers and developers, and aims to assist them when evaluating the system they have developed. It is an in-house evaluation of the technical level the system has achieved to date, and whether it has met its developmental objectives. This set of criteria evaluates the system overall, as well as individual technical components of the system. As with the other criteria, the results from the questionnaire can be plotted on a radar chart and this gives an immediate comparison of the current state of the system and its optimal state.

The significant development with these criteria compared to previous research is that in each case the criterion can be objectively derived (objective ratings), can be assigned numerical values (numerical ratings) and can be represented visually (visual judgement) in the form of radar charts (Nomura & Ishara 1992). In previous research numerical ratings had been derived; however, the introduction of objective ratings and visual judgement is an important development in non-automatic metric based evaluation. The idea of being able to compare radar charts when deciding on the best MT system for the users' or developers' needs is very advantageous, as often a visual judgement is much more effective than a comparison of percentages. Hutchins (1993:26) is correct in saying that "this impressive publication represents probably the most important single contribution yet to the literature on MT evaluation; it must surely be essential reading for anyone concerned with the development of evaluation methodologies for MT systems". We should also note here, however, that the scope of the JEIDA publication is very broad, as it takes into account many elements of an evaluation of an MT system, and their proposed evaluation methodologies would need to be tailored to the needs of evaluators in individual cases.

The report concludes with an outline of the authors' ideas for the evaluation of MT quality. The proposed approach has two aims: firstly for the developer to assess what linguistic phenomena the system cannot deal with, and secondly for the user to assess whether or not the chosen system can deal with linguistic phenomena that must be translated.

The idea of gathering opinions from both the developers and users was a novel step in MT evaluation, as was the introduction of questionnaires to gather this information.

#### 2.1.3.2 Establishing General Frameworks for Evaluation

##### EAGLES

EAGLES is the first of the three reports discussed here that deal with creating standards or a framework for the evaluation of MT. The European EAGLES initiative (Expert Advisory Group on Language Engineering Standards) is a two-phase initiative that came into being in 1993 "as an attempt to create standards for language engineering" (Hovy et al. 2002a:46) and was the first of its kind to be established. The perception at the time was that linguistic resources were essential to progress in the area of natural language processing. Therefore the aim of the initiative was to agree on standards for the form and content of resources which would make them more transferable across projects. The EAGLES group also decided to develop a general framework for evaluation design, given that it is not possible to develop one single evaluation scheme for all possible areas of NLP. The aim was to guide researchers designing their own evaluation methodology, making it easier to understand the various design decisions and compare results from different studies. The first phase of the initiative took place from 1993-1995 and the areas the initiative concentrated on were corpora, lexicons, grammar formalisms, and evaluation methodologies (Hovy et al. 2002a:46).[38] The EAGLES project took as its starting point the ISO standard 9126 and made changes to the standard to extend its scope and make it more concrete. Relevant ISO standards will be explored in more detail in the next section.

[38] The second phase of the EAGLES initiative ran from 1995-1999, with the work on evaluation limited to bringing together the guidelines and sharing these among researchers outside of the project.

### ISO Standards for Quality

We introduce International Organization for Standardization (ISO) standards at this stage, as they are related to work within EAGLES and the quality models developed for evaluation purposes. Within the EAGLES initiative, there was a group that specifically dealt with MT evaluation, namely the EAGLES Evaluation Work Group (EWG). The quality model proposed by EWG is influenced by earlier work within the group, including work based on the ISO/IEC 9126 and 14598 standards relating to software evaluation. The standards developed by the ISO ensure desirable characteristics of a product such as quality, which is the one of particular interest to the previous studies on MT evaluation and to the current study.

The objective of the particular standard used by the EWG (ISO/IEC 9126) is to "provide a framework for the evaluation of software quality".[39] According to the ISO/IEC 9126 standard, *quality* is defined as:

> The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.

ISO/IEC 9126 provides a quality model which is applicable to every kind of software, hence the use of the standard by the EWG. This standard as well as ISO/IEC 14598 has since been superseded by the SQuaRE (Software product Quality Requirements and Evaluation) framework, which follows the same general concepts.

Like quality models that preceded it (see McCall 1977), the ISO quality model classifies software quality based on three levels: characteristics, sub-characteristics and metrics. According to this model software quality results from six generic quality characteristics which are measurable attributes, and each of these has sub-characteristics associated with them, and related metrics. The six quality characteristics are *functionality*, *reliability*, *usability*, *efficiency*, *maintainability*, and *portability* (ISO/IEC 2001b). Since the first publication of the standard in 1991, updated versions have been published, mostly recently in 2001, in which the sub-characteristics have been moved from the annex to become part of the standard. Many of the original sub-characteristics have

[39] <http://www.cse.dcu.ie/essiscope/sm2/9126ref.html> [Accessed 10 March 2009].

been reworded and several new ones added. The sub-characteristics are evaluated by a set of metrics, with some metrics common to a few sub-characteristics.

According to ISO/IEC "a *measurement* is the use of a *metric*[40] to assign a value (i.e., a measure, be it a number or a category) from a scale to an attribute of an entity" (ISO/IEC, 1999, emphases original, in Hovy et al. 2002a). A metric is applied to each of the sub-characteristics in the evaluation, and then these metrics generate a measured value. There are human and automatic evaluation metrics in MT and some of these have been outlined already in previous sections. Within the EAGLES initiative the first attempts at providing a theoretical framework were put into practice via simple examples relating to language technology: quality models were developed for spelling checkers, and the initiative also looked into developing quality models for grammar checkers and translation memory systems (TEMAA, 1996 in Hovy et al. 2002a). The EAGLES initiative was a very good starting point for further research on frameworks in MT evaluation, discussed in the next two sections.

### ISLE

ISLE (International Standards for Language Engineering) is both the name of a project and the name of a group of coordinated activities within the HLT (Human Language Technology) field. ISLE, which was first proposed in 1999, acts under the auspices of EAGLES. Among ISLE's main objectives were the *development*, *dissemination* and *promotion* of de facto HLT standards and guidelines for language resources, tools and products (ISLE 1999). The EWG of the ISLE project concentrates on MT systems, refining and extending taxonomies that had been proposed by EAGLES. Within MT evaluation, the two areas of interest are building quality models for machine translation systems and maintaining and updating previous guidelines (specifically relating to ISO 9126 and ISO 14598). In order to refine its evaluation methodology and framework, the EWG in ISLE conducts a series of practical workshops that offer a forum to test out the current evaluation methodology and framework. These workshops invite both practitioners and industrialists to work on practical evaluation tasks, some of them

[40] Hovy et al. (2002:71) point out that the ISO's use of the term "metric", which Hovy et al. adopt, is not equivalent to the mathematical use of the term, as some of the scoring methods used in MT evaluation do not always have the mathematical properties of a metric. A mathematical metric has three properties, one of which relates to distance between elements. Evaluation "metrics" can sometimes be conceived as the distance between a system's response and a set of ideal responses. But depending on the scoring method used in the evaluation, not all evaluation "metrics" satisfy this property.

chosen by the EWG, while others are proposed by the participants themselves. The feedback received during and after the workshops enables the group to further develop the evaluation methodologies and the framework (Hovy et al. 2002a:60). The framework the EWG in ISLE is working on is called FEMTI (Framework for the Evaluation of Machine Translation in ISLE).

In relation to the general theory behind evaluation methodologies for HLT applications, FEMTI recognised that there is no simple answer to the question of which is the best MT system, but that the concept of a successful approach to MT evaluation is one that can be developed. They did not set out to propose new metrics, to attempt to automate the evaluation process, or to analyse the performance of human evaluators (Hovy et al. 2002a:44). In actual fact, many of the methods and metrics outlined in the Van Slype (op. cit.) report are included in FEMTI. One of the main factors mentioned by many published evaluations of MT systems is the consideration of the context of use of the MT software. This factor is not one which is given a leading role within the ISO standards. Therefore ISLE included this consideration when developing its framework, allowing an MT evaluator to determine a particular quality model based on the expected context of use.

### FEMTI

FEMTI is the aforementioned framework that takes into consideration the context of use of the MT system. The framework[41] is made up of two interrelated classifications or taxonomies, and each feature from the first classification is linked to relevant quality characteristics and metrics in the second. The first of these three elements is a classification of the main features defining a context of use in terms of the users of the MT system, the task the system is being used for, and the type of input; the second element is a classification of quality characteristics of the MT software (see Figure 2.1 overleaf), which are further broken down in sub-characteristics, and below this attributes/metrics. Examples of metrics are underlined in Figure 2.1 for *comprehensibility*, *readability*, *style* and *well-formedness*.[42] It should be noted that FEMTI does not provide metrics for all sub-characteristics. Finally, the third element

[41] Fiederer & O'Brien (2009:55) point out that "MT evaluation is hampered by the use of synonyms to describe evaluation parameters and the creation of FEMTI was an attempt to gather into one place the accumulated experience of MT evaluation."
[42] We have highlighted these metrics as they are the four metrics we measure in the present study.

maps from the first to the second classification, helping to define or suggest quality characteristics and associated metrics that are considered the most relevant for each context of use. It is envisaged that MT evaluators will be able to adapt this framework according to the nature of context of use of the particular MT software or a particular element of the software they are evaluating.

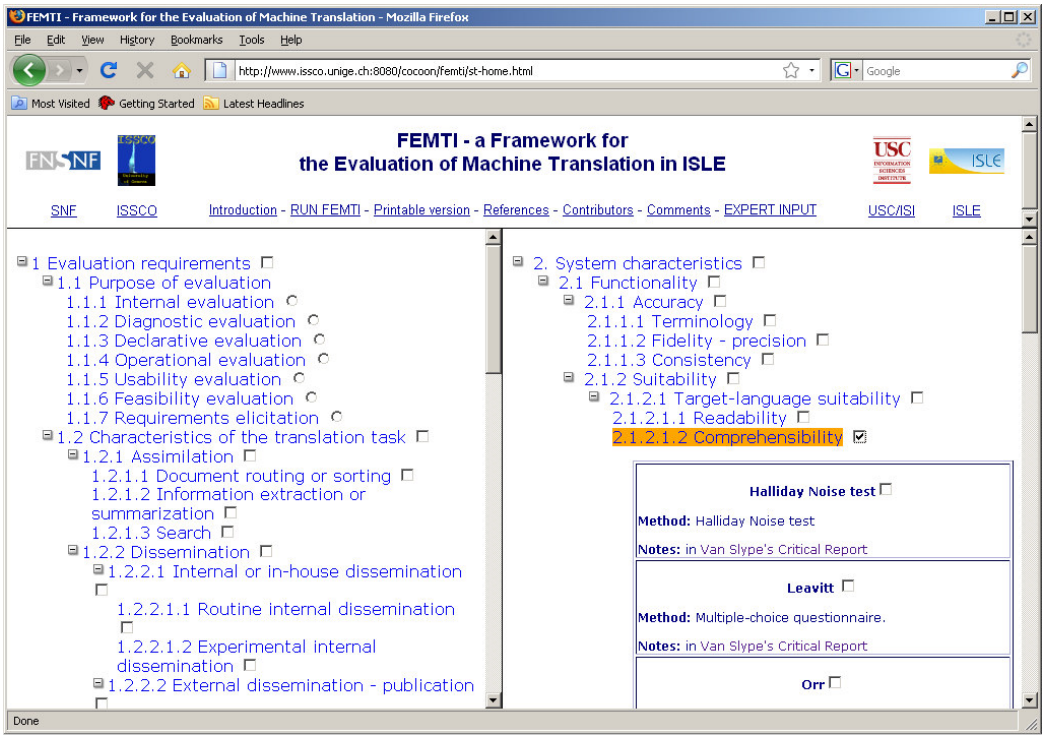
**Figure 2.1:** FEMTI classification of quality characteristics of MT software provided in the second taxonomy

**2.1 Functionality**
- *2.1.1 Accuracy*
  - 2.1.1.1 Terminology
  - 2.1.1.2 Fidelity/precision
  - 2.1.1.3 Well-formedness
    - 2.1.1.3.1 Morphology
    - 2.1.1.3.2 Punctuation errors
    - 2.1.1.3.3 Lexis/Lexical choice
    - 2.1.1.3.4 Grammar/Syntax
  - 2.1.1.4 Consistency
- *2.1.2 Suitability*
  - 2.1.2.1 Target-language suitability
    - 2.1.2.1.1 Readability: Cloze test; Subjective rating of intelligibility; Reading time
    - 2.1.2.1.2 Comprehensibility: Halliday noise test; Leavitt multiple-choice questionnaire
    - 2.1.2.1.3 Coherence
    - 2.1.2.1.4 Cohesion
  - 2.1.2.2 Cross-language/Contrastive
    - 2.1.2.2.1 Style: Van Slype evaluation of sentences; String-edit distance
    - 2.1.2.2.2 Coverage of corpus-specific phenomena
  - 2.1.2.3 Translation process models
    - 2.1.2.3.1 Methodology
      - 2.1.2.3.1.1 Rule-based models
      - 2.1.2.3.1.2 Statistically-based models
      - 2.1.2.3.1.3 Example-based models
      - 2.1.2.3.1.4 TM incorporated
    - 2.1.2.3.2 MT Models
      - 2.1.2.3.2.1 Direct MT
      - 2.1.2.3.2.2 Transfer-based MT
      - 2.1.2.3.2.3 Interlingua-based MT
  - 2.1.2.4 Linguistic resources and utilities
    - 2.1.2.4.1 Languages
    - 2.1.2.4.2 Dictionaries
      - 2.1.2.4.3 Word lists or glossaries
      - 2.1.2.4.4 Corpora
      - 2.1.2.4.5 Grammars
    - 2.1.2.5 Characteristics of process flow
      - 2.1.2.5.1 Translation preparation activities
      - 2.1.2.5.2 Post-translation activities
      - 2.1.2.5.3 Interactive translation activities
      - 2.1.2.5.4 Dictionary updating
- *2.1.3 Well-formedness:* Percentage of phenomena correctly treated; Average string-edit distance per sentence or for all inflectable tokens in the text; List of error types
- *2.1.4 Interoperability*
- *2.1.4 Functionality compliance*
- *2.1.5 Security*

**2.2 Reliability**
*2.2.1 Maturity*
*2.2.2 Fault tolerance*
*2.2.3 Crashing frequency*
*2.2.4 Recoverability*
*2.2.5 Reliability compliance*
**2.3 Usability**
*2.3.1 Understandability*
*2.3.2 Learnability*
*2.3.3 Operability*
2.3.3.1 Process management
*2.3.4 Documentation*
*2.3.5 Attractiveness*
*2.3.6 Usability compliance*
**2.4 Efficiency**
*2.4.1 Time behaviour*
2.4.1.1 Overall Production Time
2.4.1.2 Pre-processing time
2.4.1.3 Input to Output Translation speed
2.4.1.4 Post-processing time
2.4.1.4.1 Post-editing time
2.4.1.4.2 Code set conversion
2.4.1.4.3 Update time
*2.4.2 Resource utilisation*
2.4.2.1 Memory usage
2.4.2.2 Lexicon size
2.4.2.3 Intermediate file clean-up
2.4.2.4 Program size
**2.5 Maintainability**
*2.5.1 Analysability*
*2.5.2 Changeability*
2.5.2.1 Ease of upgrading multilingual aspects
2.5.2.2 Improvability
2.5.2.3 Ease of dictionary update
2.5.2.4 Ease of modifying grammar rules
2.5.2.5 Ease of importing data
*2.5.3 Stability*
*2.5.4 Testability*
*2.5.5 Maintainability compliance*
**2.6 Portability**
*2.6.1 Adaptability*
*2.6.2 Installability*
*2.6.3 Portability compliance*
*2.6.4 Replaceability*
*2.6.5 Co-existence*
**2.7 Cost**
*2.7.1 Introduction cost*
*2.7.2 Maintenance cost*
*2.7.3 Other costs*

The FEMTI resource is available online, written in XML format, which allows the user to interact with the framework. The evaluator is able to specify the intended context of use from the first taxonomy (see Figure 2.2, left-hand side) and submit this to FEMTI. Following this, FEMTI will generate a proposed set of quality characteristics that it considers relevant to that context. All of the suggestions made by FEMTI are highlighted in yellow (see Figure 2.2, right-hand side). If the evaluator wants to amend this list and add any new quality characteristics, these are highlighted in orange. If the evaluator wants to remove any of the characteristics suggested by FEMTI within the second classification, they simply uncheck the box (see Figure 2.2, right-hand side). FEMTI is a continually evolving resource that encourages its users to offer feedback, which it then incorporates on a regular basis. This framework is intended to help various groups of people, one of them being those who want to evaluate the suitability of a given system with respect to a given task, which is one of the aims of the current research. Therefore FEMTI played an important role in the development of the proposed evaluation methodology outlined later in this chapter, and expanded on in Chapter 3.

**Figure 2.2:** Online FEMTI resource. The first taxonomy on the left outlines the evaluation type and the context characteristics. The second taxonomy on the right outlines the quality characteristics, sub-characteristics and associated metrics



#### 2.1.3.3 Automatic Evaluation of Translation Quality

##### Move towards automatic metrics

Over the last decade there has been a move away from rule-based MT systems and now most MT research is based on corpus-driven systems, as outlined in Chapter 1. This move towards corpus-driven approaches is coupled with a move from human evaluation to automatic evaluation methods. Papineni et al. (2002) describe automatic metrics as a cheap way to generate and provide results within a matter of seconds in many cases. Since it is claimed by the developers of automatic metrics that they correlate well with results from human judgements of MT output (cf. Papineni et al. 2002, Doddington 2002, Coughlin 2003, Banerjee & Lavie 2005, Snover et al. 2006), we could possibly use these metrics to estimate the potential performance of the system based on the scores generated. These metrics can also indicate whether or not a system is improving, as higher scores in consecutive evaluations indicate that the output generated is more similar to a 'gold standard' human reference translation, and this output correlates well with human judgements (from previous trials of the metric). Automatic metrics are thus useful for carrying out interval testing while developing a system, and when researchers wish to find out if certain changes are reflected in the scores produced by metrics.

##### BLEU, NIST, WER, PER, Precision and Recall

In 2002 Papineni et al. devised a method for automatic evaluation of MT, calling it the **bil**ingual **e**valuation **u**nderstudy, or BLEU. Their motivation for developing this method resulted from the high costs of and long amounts of time needed for human evaluation. They described this metric as an "automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations" (Papineni et al. 2002:311). This is probably the most well-known automatic metric for evaluating MT output. The aim of BLEU is to estimate the closeness of the translation output to a reference or 'gold standard' translation(s), and act as an inexpensive automatic evaluation "that is quick, language-independent, and correlates highly with human evaluation" (ibid).

BLEU is based on the *word-error rate* (WER) metric used within the speech recognition community. WER is a metric based on edit-distance between two strings and it is often referred to as the Levenshtein distance. WER is a measure of the number of modification operations required to transform one sentence (output) to a standard

human translation in terms of the number of insertions, deletions and substitutions required. Another metric closely related to WER is *position-independent word-error rate* (PER); the only difference is that we ignore the position of the words in the sentence when calculating PER. With WER and PER metrics, the lower the score, the better.

The central idea behind BLEU is that *the closer a machine translation is to a professional human translation, the better it is* (ibid). The BLEU metric thus rewards candidate translations (MT system output) which are most similar to the reference translations. When calculating BLEU two scores must be considered: *n*-gram precision ($p_n$) and the *brevity penalty* (BP). *N*-gram precision is related to the number of sequences of words (1-gram = 1 word, 2-gram = 2 words etc.) the candidate translation has in common with the reference translation(s). The process can in theory be applied to any number of *n,* but in practice *n* is capped at four.

The brevity penalty relates to the length of the candidate sentences. Papineni et al. (ibid) state that a candidate sentence should be neither too long nor too short, and the evaluation metric should enforce this. To some extent the *n*-gram precision does just that. It penalizes unauthentic or contrived words in the candidate translations. However, it cannot enforce proper translation length on its own. This is where the BP steps in. By using the BP, a high-scoring candidate translation must match the reference translation(s) in length, in word choice, and in word order. When the length of the candidate translation is the same as or longer than the reference translation, the BP is 1.

To calculate the BLEU score the geometric mean[43] is taken of the precision scores for each segment and the result is multiplied by an exponential BP factor (see Papineni et al. 2002 for a more detailed explanation of the formulae mentioned here). An important feature of BLEU is that it is a text-based metric, developed with document or system level evaluation in mind, and does not focus on generating scores that correlate well with human judgements at sentence level. Papineni et al. say that BLEU only needs to match human judgement when averaged over a test corpus, given that scores on individual sentences will often vary from human judgements. Many researchers argue against this (cf. Kulesza & Shieber 2004, Liu & Gildea 2005, Banerjee & Lavie 2005,

[43] The geometric mean is calculated by multiplying the individual values in a set and taking the nth root of the product.

Amigó et al. 2006, Owczarzak 2008) calling for more focus on reliable, automatic, sentence-based metrics, as text-based metrics might not be suitable in all contexts. Doddington (2002) and Banerjee & Lavie (ibid) are both critical of BLEU's use of the geometric mean. Banerjee & Lavie (ibid:67) explain that if one of the component *n*-gram scores is zero, then using the geometric mean will give an average score of zero. Therefore, despite a candidate translation receiving relatively high BLEU scores for 1-gram and 2-gram counts, the final score would be zero if the 3-gram count was zero. Liu & Gildea (ibid:26) and Owczarzak (ibid:4) suggest the need to move away from using *n*-gram sequences altogether, saying that identifying word matches is not a satisfactory way of measuring translation quality, and automatic metrics should aim at recognising well-formed and more readable sequences, rather than simply matching words.

Another metric, very similar to BLEU, is NIST (National Institute of Standards and Technology,[44] Doddington 2002). The only difference between the two metrics concerns the calculation of n-gram co-occurrences: BLEU uses the geometric mean to calculate the precision scores, whereas NIST uses the arithmetic mean[45] to calculate the same scores. In contrast to WER and PER, both BLEU and NIST scores range from 0 to 1, and show the statistical closeness of the output to one or more reference translations. The higher the score the better the quality of the translation.

Two other automatic metrics that can be used in relation to MT evaluation are Precision and Recall (Turian et al. 2003). Precision and recall often have an inverse relationship: by increasing the precision, the recall can be reduced and vice versa. F-measure is related to both precision and recall, as it is the weighted harmonic mean of the two:

$$\text{F-measure}_{\text{alpha}} = \frac{(1 + \text{alpha}) \times \text{precision} \times \text{recall}}{((\text{alpha} * \text{precision}) + \text{recall})}$$

From the formula, as the alpha value increases, the weight of recall increases in the measure. Turian et al.'s metric, *General Text Matcher* (GTM) allows the use of multiple reference translations to calculate precision, recall and f-measure by calculating the

[44] <http://www.nist.gov/> [Accessed 10 March 2009].
[45] The arithmetic mean is the sum of the individual values in a set divided by the number of values in the set.

overlap between candidate and reference translations as the maximum subset of non-repeated words present in both texts. GTM also assigns a higher weight to longer matches and to matches in the right order (Estrella 2008:37). All of the abovementioned metrics are used to automatically evaluate MT output, and to rank MT systems according to the results.

For many researchers the task of choosing between human evaluation methods and automatic evaluation methods boils down to the cost of and time involved in implementing either one. As already indicated, human evaluations are said to be very time consuming and extremely costly, and in many respects this is true. Papineni et al. (2002:311) mention that human evaluations can take weeks or months to finish, and that this is not satisfactory when developers need to monitor the effect of daily changes to their systems "in order to weed out bad ideas from good ideas." They also mention that automatic metrics can be used when there is a "need for quick or frequent evaluations" (ibid). In this respect automatic evaluation metrics are indeed very useful. The automatic metrics mentioned have been shown to correlate well with human judgements as outlined in Papineni et al. 2002 (BLEU), Doddington 2002 (NIST), Turian et al. 2003 (precision, recall, F-measure) and Lavie et al. 2004 (precision, recall, F-measure), which means BLEU and NIST have remained popular among MT researchers, who wish to compare the relative quality of different MT outputs. A good example of this is the use of these metrics in organised MT shared tasks[46] to evaluate machine translation performance (Koehn & Monz 2006). More recently, however, researchers have highlighted that although BLEU and NIST measures are useful for comparing the quality of outputs, they believe it is difficult to see what the scores actually mean and that both BLEU and NIST do not correlate well with human judgement scores at sentence or segment level, even if they do correlate well at paragraph or text level (Turian et al. 2003, Snover et al. 2006, Lavie & Agarwal 2007, Callison-Burch et al. 2007, Volk 2008). This inadequate accuracy of evaluation at segment level has been widely criticised, for example in Callison-Burch et al. (2006), who showed that BLEU failed to recognise allowable variation in translations in cases where multiple reference translations were not available. In the following Example 2.1 (A and B) taken from Owczarzak (2008:15), we can see that the score in example A is artificially lowered due

[46] Within the SMT community it is common practice to hold a shared task, the aim of which is to evaluate MT systems using a pre-defined data set between groups who apply to participate.

to the absence of a 4-gram in common with the reference(s) and the example in B contains less than four elements, meaning it will be scored as zero, irrespective of lower n-gram matches in the sentence. In contrast, human evaluation of these sentences would result in a perfect score.

**Example 2.1:** Calculating automatic metric scores using reference translations

(A) **Translation**: John resigned from his job yesterday.
**Reference**: Yesterday John quit his job.

(B) **Translation**: John resigned.
**Reference**: John quit.

Callison-Burch et al. (2006) and Owczarzak et al. (2006) comment on instances where *n*-gram based metrics such as BLEU and NIST were shown to be biased towards statistical MT output that makes use of a statistical-based decoder, such as Pharaoh (Koehn 2004), and output generated by rule-based systems was consistently ranked lower using automatic metrics, contradicting the human judgements of the same output.

We can see from these examples above why the BLEU metric might not be optimal for use with subtitles, given that subtitles are often short segments, as in Example 2.1, (B). Several other proposed metrics including METEOR (Lavie & Agarwal 2005, 2007), GTM (Melamed et al. 2003), TER (Snover et al. 2006) and CDER (Leusch et al. 2006) aim to address some of the weaknesses associated with correlations with human judgements at sentence/segment level. Of these metrics our system generates METEOR scores (see Chapter 5).

In addition to ranking systems based on automatic metrics, shared MT tasks also incorporate human evaluation of the output. Human evaluation results are usually taken to be authoritative and are then used to evaluate the automatic metrics in "meta-evaluation" (Callison-Burch et al. 2008). The number of human evaluators involved can vary depending on the number of participating groups in the task and the number of other volunteers willing to take part. In the 2007 ACL shared task, over 100 people participated in the evaluation, which was nearly double the number who participated in the 2006 human evaluation. Usually texts from the Europarl Corpus are used in the evaluation, and in 2007 news editorials were also added into the test set. Within the MT

research community, the most commonly used human evaluation methodology is to assign values from two 5-point interval scales, representing fluency and adequacy. These scales were developed by the Linguistic Data Consortium (LDC 2005) and are shown in Table 2.2 below.

**Table 2.2**: Scales for adequacy and fluency developed by LDC (2005)

| **Adequacy**<br>How much of the meaning in the HT reference is expressed in the MT output? | **Fluency**<br>How fluent is the MT output? |
|---|---|
| 5 = All<br>4 = Most<br>3 = Much<br>2 = Little<br>1 = None | 5 = Flawless English<br>4 = Good English<br>3 = Non-native English<br>2 = Disfluent English<br>1 = Incomprehensible |

These scales are more comparable with those suggested by Van Slype's (1979) 4-point scale and the 5-point scales used during the DARPA evaluations, and they make fewer distinctions within a translation compared to the scales suggested by Pierce et al. (1966). The simplification of rating scales seems to coincide with the overall simplification of the human evaluation design implemented in the shared tasks. In the ACL 2007 shared task (Callison-Burch et al. 2008), only 40% of the MT output was evaluated by more than one evaluator, and in the IWSLT[47] 2006 evaluation (Paul 2006), only the Chinese-English translations were selected for human evaluation. These changes to the evaluation design are understandable given the large amounts of data generated,[48] costs involved in conducting a thorough human evaluation of every output sentence and the hours of labour invested in the tasks.[49] However, we must examine if the simplification of the human evaluation design has had a negative effect on the reliability of the results. Callison-Burch et al. (2007) note that in the evaluation the

[47] International Workshop on Spoken Language Translation.
[48] At the ACL 2007 task there were 15 participating systems, a slight increase on the ACL workshop from the previous year, and the IWSLT 2006 received 21 submissions from 19 research groups.
[49] In 2006, the manual evaluation was conducted by volunteers, and amounted to nearly 180 hours of labour. The 2007 task saw a collective total of 330 hours of labour, an increase of almost 150 hours on the 2006 task, and there were also a small number of paid annotators in addition to the larger number of volunteers. The task in 2008 followed the same pattern with a small number of paid annotators, but saw a reduction of nearly 70 hours of labour on the 2007 task.

observed Kappa[50] coefficient value for judgements of fluency was 0.25 and for judgements of adequacy was 0.226; intra-annotator agreement (i.e. an average measure of the evaluator's consistency) produced a Kappa coefficient value of 0.537 for judgements of fluency and 0.468 for judgements of adequacy. Similar low results were detected in the IWSLT 2006 task, in which there were Kappa values of 0.24 for fluency and 0.31 for adequacy. These Kappa values were deemed to indicate unreliable fluency and adequacy judgements, and called to question the role of human evaluations as gold-standards. This led the ACL 2008 shared task (Callison-Burch et al. 2008) to reconsider the design of the manual evaluations, producing a refined design that required the subjects to evaluate the MT output in three different ways, with the main difference from the previous year being the removal of the fluency and adequacy scales to rate individual sentences, and the inclusion of a yes/no judgement on the acceptability of syntactic constituent translations in order to increase inter- and intra-annotator agreement. Two of the three evaluation tasks made judgements on shorter segments than tasks in previous years. The results showed *fair* inter-annotator agreement and *moderate* intra-annotator agreement for ranking translations of whole sentences relative to each other. Inter-annotator agreement Kappa scores for ranking translations of syntactic constituents taken from the source sentences were *moderate*; and for assigning yes/no judgements to the acceptability of these translations were substantial. Intra-annotator agreement was substantial to almost perfect for both of these tasks. These results from the tasks involving shorter phrases restore some confidence in manual evaluation, and suggest that perhaps the exercise of ranking whole sentences is simply too complex, given that sentences can be ranked according to numerous sentence characteristics.

The MT (human) evaluation studies surveyed here all follow a similar pattern of using scales to rate a quality characteristic of a sentence to decide whether the output meets a required standard. Over time these human evaluations have been simplified, with a possible negative effect on their reliability. However, the reliability of manual evaluations has been investigated, resulting in a new design and new scales. We have seen that human evaluation strategies can be costly and time-consuming, which

[50] Cohen's Kappa (often referred to as Kappa or Kappa coefficient) is a measure of agreement between individuals rating the same thing. Landis & Koch (1977) interpret Kappa values as 0-0.2 indicating slight agreement, 0.21-0.4 indicating fair agreement, 0.41-0.6 indicating moderate agreement, 0.61-0.8 indicating substantial agreement, and above 0.8 indicating almost perfect agreement.

encouraged a growth in the use of automatic metrics. As mentioned previously in the Introduction, automatic metrics have been described as an 'imperfect substitute' for human evaluation of translation quality, and therefore in the ACL shared tasks, the manual evaluation takes priority, and the human judgements validate the automatic metrics. Using human judgements in this way increases the reliability of automatic metrics, which can then be applied to new areas of translation quality research.

In the next section we introduce the human evaluation model devised for this study, and describe how we incorporate previously tested human evaluation methods and automatic metrics into the study, and combine them with methods not previously used for human MT evaluation, but which are appropriate in the context of subtitling.

## 2.2 MT Evaluation in the Current Study

There are important choices to make when designing any evaluation methodology. One of the most important is whether to choose human evaluation methods or automatic evaluation methods, or indeed a mixture of both. We have already examined the kinds of human and automatic evaluation approaches commonly used in an MT research environment, and we identified the strengths and weaknesses of both. Chapter 1 highlighted the different approaches to evaluating machine-generated subtitles including one translator making judgements on fidelity and grammaticality much like the studies outlined above (Popowich et al. 2000), while the remaining studies calculated BLEU scores to evaluate quality (Melero et al. 2006, Piperidis et al. 2005 and Volk 2008). Armstrong et al.'s (2006c) pilot study used a series of different manual evaluation strategies to investigate which one produced the most useful data to answer questions related to end-user acceptability. Their pilot study highlighted how the commonly used method for human evaluation of MT, which involves assigning subjective scores to the translation of individual sentences in a text-based context, could not be applied to the context of subtitling. According to Gottlieb (2005:19) there are four basic semiotic channels used in filmic media: image, writing (including displays and captions), sound effects (including music and effects added in post-production) and speech (excluding inaudible background dialogue). In the case of subtitling, he believes the relative impact of semiotic channels can be divided up in the following way (based partly on his personal experience as a subtitler, and partly on theoretical studies cf. Gottlieb 1997):

**Table 2.3**: Relative impact of semiotic channels in subtitling

| **Subtitling** | |
|---|---|
| Image | 40% |
| Writing | 32% |
| Sound effects | 18% |
| Speech | 10% |

Table 2.3 claims that approximately 32% of the semantic load is communicated to the target audience through writing on the screen, which during a subtitled movie would usually be subtitles, as opposed to displays or credits. This claim strengthens the argument that subtitles should be evaluated only in a 'natural' setting, where other semiotic channels are also available, and that viewers construct meaning based on all present semiotic signs, verbal and nonverbal.

The following sections describe the two-phase approach to human evaluation of EBMT subtitles adopted in this research.

### 2.2.1 Developing the Evaluation Model: The Two-Phase Approach

The evaluation strategy adopted in this study is primarily a human evaluation. In keeping with current research in the MT community, we also calculate scores using automatic metrics for our system output. However, the main focus is on manual corpus analysis and human judgements on machine-translated subtitles.

The evaluation is divided into two phases: prospective and retrospective. The aim of the prospective phase is two-fold: to analyse the three corpora used in this study in terms of specific characteristics (e.g. their repetitiveness), and to investigate if instances of repetition in a corpus are related to judgements on the reusability of the corresponding translations in different contexts.

The prospective phase is essentially a corpus-analysis study. We firstly wanted to (compile and) evaluate three corpora (A, B and C) for use with the EBMT system. During this phase we implement Volk's (2008) corpus 'profiling', as outlined in section 1.5.5. The three corpora contain subtitles taken from commercial DVDs. Before using the subtitles to compile the corpora, we took a sample of the subtitles and conducted a

recipient evaluation with three German-native speakers to judge the quality of the chosen sample as an indication of the overall quality of the corpus (see Appendix A for the results of this evaluation). Once our corpus was compiled we carried out an analysis of the data it contains.

If we refer to the ACL shared tasks or indeed the IWSLT conferences over the past few years, the training data are supplied to the system developers and the only data statistics supplied are the number of sentences, the number of words and the number of distinct words (type/token ratios). The corpora were not analysed in advance of training the system to identify characteristics that may have an impact on the quality of the output. The 2008 ACL shared task introduced a new test set consisting of news commentaries (described as out-of-domain texts) to see if some systems would perform better when translating texts from a different domain to that of the training data, and this change did favour some rule-based systems. However, these improvements were identified *a posteriori* on the basis of improved automatic metric scores for the MT output.

The three corpora in the current study are analysed based on the following characteristics: the number of SL repetitions in the corpus and the possible reusability of the corresponding TL translations in new contexts; we also take into account the size of the corpora and the homogeneity (genre). If, in advance, these characteristics could be identified as contributing factors to the improved quality of MT output (in terms of intelligibility and acceptability), this would be very beneficial to system developers.

EBMT systems work on the basis of using previously stored examples to produce MT output. From the point of view of an EBMT system, we needed to establish whether or not we could assume that detecting high levels of repetition in the corpus before we use it with the EBMT system would mean better recall of examples at segment and sub-segment level[51] at EBMT run-time. Of course as we have noted before, high levels of recall do not necessarily mean high levels of precision, something which affects the acceptability of the subtitles as assessed by the end-user. In contrast, if the corpus does not contain any repetitions or at least displays a very low level of repetition at segment and sub-segment level, we would assume that the low level of recall would have a negative effect on the acceptability of the subtitles as it would be more difficult for an

[51] Sub-segments can be analysed using the Trados tool if the corpus is segmented at sub-segment level (See section 4.1).

EBMT system or indeed a hybrid EBMT/SMT system such as the one used in our study to generate output based on previous 'good' examples. We describe the EBMT system in detail in Chapter 3; however, it suffices to say at this stage that the EBMT system can store any aligned examples in the training corpus, so that these examples can be re-used during the translation process. Therefore in the case where a corpus does not contain any segment or sub-segment repetition, or indeed has low levels of repetition, the corpus would have to be supplemented with a large amount of training data to build up its levels of repetition. This of course is costly and time-consuming, and often not feasible in many cases. Once we establish the repetition levels in the corpora, we investigate if instances of repetition in a corpus are related to the reusability of translations in different contexts. For instance, if the English subtitle 'Come on' has already been translated as 'Komm schon', we are interested in whether or not 'Komm schon' can be used to translate other instances of 'Come on' subsequently encountered in a particular movie or group of movies.

The retrospective phase, on the other hand, employs a context-based declarative evaluation derived from the FEMTI[52] model and a recipient evaluation as outlined by Trujillo (1999). The approach consists of end-user evaluation sessions, in which subjects view movie clips with subtitles, and the researcher then administers a questionnaire during a retrospective interview. Machine-generated subtitles are presented to the subjects in an 'experimental' real-world setting.[53] The aim of the end-user sessions is to collate the opinions of the end-users of subtitles using a combination of a retrospective interview and questionnaire. Both methods aim to investigate particular quality characteristics of the EBMT-generated subtitles, and we use the data thus elicited to establish the intelligibility and acceptability of the subtitles. In this second phase we want to also validate or refute any assumptions we might have made regarding repetition and reusability during the first phase.

Currently, we evaluate the output from the point of view of the end-user of the output and not from the point of view of the end-user of the MT system, i.e. the subtitler. We are not interested in the usability of the interface of the system, as the viewers of the

[52] FEMTI bases its definition of a declarative evaluation on the work of White (2000). White's work on the types of evaluation methodologies was published in 2000 and 2003. The work we have referenced in this study is from his 2003 publication. FEMTI reference the work from his 2000 publication. However, the same information on evaluation methodologies is provided in both.
[53] The details of this setting are elaborated on in Chapter 3.

DVD clips do not have any access to the interface. We are also not concerned with the speed of the system; however, we may note here that it takes only a few minutes to generate output using the type of MT system in this study. Therefore the scope of the evaluation includes only the MT output from the end-users' perspective. The context of use of the output is German subtitles for a movie on DVD.

### 2.2.2 Quality Characteristics in this Study

We drafted the declarative element of the overall evaluation model for this study using the online FEMTI resource. A context-based evaluation model consists of the intended context of use and a quality model (quality characteristics and metrics). The first step involves choosing relevant characteristics for the context of use. For this study we chose a declarative evaluation, and the context characteristics included:

- communication,
- domain or field of application (DVD subtitles)
- genre (fantasy subtitles)
- document type (subtitles)
- synchronous communication
- characteristics of the translation task
- characteristics related to sources of error

These evaluation requirements were then submitted to the FEMTI tool, and FEMTI displayed the quality characteristics relevant to the selected context of use. Given that this study aims to evaluate the intelligibility and acceptability of machine-generated subtitles, we chose only the characteristics related to these two concepts. Table 2.4 outlines the relevant quality characteristics suggested by FEMTI, including suitable metrics to measure the data:

**Table 2.4:** Quality characteristics and associated metrics suggested by FEMTI

| Quality Characteristic | Metrics |
|---|---|
| Comprehensibility | • Halliday noise test<br>• Sinaiko knowledge test<br>• Orr & Small multiple-choice questionnaire<br>• Leavitt multiple-choice questionnaire |
| Readability | • Crook & Bishop cloze tests<br>• Sinaiko multiple-choice questionnaire<br>• Van Slype evaluation of sentences on a 4-point scale |
| Style | • String-edit distance<br>• Van Slype evaluation of sentences on a 4-point scale |
| Well-formedness | • Percentage of phenomena correctly treated<br>• Average string-edit distance per sentence or for all inflectable tokens in the text<br>• List of error types |

### Four Quality Characteristics

We mentioned briefly in the introduction how we define the quality characteristics for this study, and we elaborate on these definitions in the next few paragraphs and discuss the methods used to measure them. It has been noted in the literature that quality is an elusive notion in relation to translation and MT in particular, and that the only way of actually measuring translation quality is to measure whether a translation is good enough for a specific purpose (context) using a combination of different criteria (King 1997, Van Slype 1979). In this study we are measuring MT output quality on the basis of intelligibility and acceptability from end-users' perspectives. Intelligibility is a necessary condition for acceptability, but it is not a sufficient condition on its own. We have identified the following FEMTI quality characteristics as being of interest to our study: comprehensibility, readability, style and well-formedness. In order to measure intelligibility we measure the comprehensibility and readability of the subtitles; in addition to this, we measure the style and well-formedness of the subtitles to ascertain

their acceptability. We measure these characteristics using qualitative and quantitative methods, some of which are suggested by FEMTI. These combined measures allow us to investigate the overall intelligibility and acceptability of the subtitles by the end-users.

In this research we understand *comprehensibility* as "the extent to which the text as a whole is easy to understand, that is, the extent to which valid information and inferences can be drawn from different parts of the same document" (Van Slype 1979:62). FEMTI suggest four methods to measure this concept, all of which are mentioned in Van Slype (1979): noise test (Halliday), two multiple-choice questionnaires (Leavitt 1971, Orr & Small 1967) and a knowledge test (Sinaiko 1978). We do not use any of these methods directly as the design is not appropriate for the context of subtitling. However, we use methods based on a combination of the methods suggested here. We administer a combined questionnaire and interview, and in the questionnaire we include a comprehension test, asking the subjects if they understood the movie clip using the subtitles and two additional questions to test their knowledge (building on the multiple-choice questionnaire and knowledge test), and we ask the subjects to rate the subtitles on a comprehensibility scale.

We understand *readability* in this study as being able to read the text quickly within a restricted time-frame, to understand the text clearly and to persevere in reading the text (Klare 1977 cited in Cadwell 2008:12). We noted earlier that some quality characteristics are used synonymously. Comprehensibility and readability have been used by authors to refer to intelligibility, clarity and fidelity (Hutchins & Somers 1992, Van Slype 1979, Pierce et al. 1966). However, comprehensibility and readability refer to different, albeit closely related concepts. Comprehensibility is defined for reading text in isolation, so that the evaluator would have a few chances to read and re-read the text to see if they understood the meaning, whereas a readability measure is applied when the evaluator only has one opportunity to read the text, and from this they must rate it. The element of a reader being able to read a text quickly differentiates readability from comprehensibility. As already indicated, viewers of subtitles usually do not get a second opportunity to re-read the subtitle as it only stays on the screen for a prescribed time. If it is only a short subtitle, it may be possible to re-read it. That said, one-line subtitles are timed to stay just long enough to avoid the viewer from starting to re-read

it. However, if the subtitle spans two lines, there would probably not be enough time to re-read the subtitle, and if a viewer tries to do so, this may have a negative knock-on effect whereby they would miss the beginning of the next subtitle making the viewing of the subtitles not very enjoyable. Metrics that allow for the re-reading of MT output are thus not appropriate to evaluation in our context.

Some methods suggested by FEMTI to measure readability include Cloze tests (Crook & Bishop 1965, cited in Halliday & Briss 1977[54]), multiple-choice questionnaires (Sinaiko 1978, cited in Van Slype 1979) and rating scales (Carroll 1966, cited in Pierce et al. 1966). The current study measures readability using a rating scale as suggested by Carroll (1966). The subjects rate how satisfied they were with the subtitles (relating to understanding of the subtitles and being able to read them in the time-frame) and we asked open questions in the questionnaire relating to the perceived speed of the subtitles (reading time) and regarding subtitles that are deemed by the viewer to be out of context (which might have caused the subjects to have trouble reading and understanding the subtitles). Cloze tests and multiple-choice questionnaires were not appropriate for measuring readability in an AVT context.

Popowich et al. (2000) maintain that typically when evaluating the acceptability of a translation system or system output, some metric *m* is chosen and then the acceptability of the system is a direct function F(*m*), where an increase in the value of *m* indicates an increase in the acceptability of the system. However a definition of this type is inadequate when dealing with time-constrained applications such as the translation of DVD subtitles. In this type of domain the acceptability of a translation depends not only on the understanding of the translation, but also the time required for its comprehension. Therefore we must take time into account when evaluating acceptability, and use a function F(*m*, *t*). Popowich et al. (ibid) comment that the effect *t* can only be calculated in an operational context. This allows the evaluator of the subtitles to judge whether or not a translation of the subtitle is acceptable in a real-world situation, and not simply text read in isolation. This measure of acceptability considers the measure of comprehensibility and readability as outlined above.

[54] According to Somers & Wild (2000) extensive efforts to obtain a copy of Crook & Bishop's report have revealed that the original was probably lost in a fire at Tufts University, so we can only go on Halliday & Briss's summary.

*Style* is defined in this study as the extent to which the translation uses the language appropriate to its content and intention (Hutchins & Somers 1992:163). Fiederer & O'Brien (2009:56) point out that while style is considered important in the rating of human translation, in particular in literary domains, it rarely figures as an evaluation parameter in the rating of MT output. Subtitle texts have been compared with literary texts in the literature (Volk 2008:7), making style an important evaluation criterion in this study. FEMTI suggests Van Slype's (1979) evaluation of sentences on a 4-point scale and string-edit distance developed by Niessen et al. (2000). We adopt Van Slype's method and ask subjects to measure style on a 6-point scale. We also use two open questions asking the subjects if something about the subtitles either bothered or amused them in relation to the appropriateness of the language. DiMarco (1994:32) points out that:

> Style influences translations on all linguistic levels: lexical, syntactic, semantic and textual. The particular choices of words, the specific arrangement of sentences, the selection of which details to carry over in the translation, all convey a particular stylistic effect.

Style is important for the present study as our target language, German, distinguishes for example, between formal and informal forms of address (cf. 'du' (you informal) or 'Sie' (you formal)). DiMarco & Hirst (1990) observe that the style used in the translation must be appropriate and natural to the target language, and this can only truly be measured by native speakers of the target language.

Style, as used in this study, is synonymous with register. Díaz Cintas & Remael (2007:189) describe register when used in audiovisual contexts as:

> [T]he concept used to denote the language produced by a particular social situation and characterized by the different degrees of formality linked to that situation.

We use the term style as it is the term used in the FEMTI model (which was previously called 'style/register') and we are measuring the quality characteristic with regard to MT evaluation. We evaluate the style of the MT output in terms of appropriateness and naturalness in the given context. We acknowledge, however, that style in the context of AVT studies concerns the manner of speech, expressions used and characterisation, and human translators usually use "compensation" strategies to address stylistic issues

(ibid:188). The quality characteristic of style, as defined within AVT, could be investigated in future studies. This means that the style questions used in the current interview questionnaire would need to be reformulated to elicit responses relating to the style of the subtitles from an AVT perspective.

Lastly *well-formedness* is defined in this study as the degree to which the output respects the reference rules of the target language at the specified linguistic level (Flanagan 1994, Arnold et al. 2003 and Loffler-Laurian 1983). FEMTI's category of well-formedness includes what they consider the four most critical categories of error typically made by MT systems. The well-formedness criterion we incorporated in this study investigates errors and the severity of these errors as judged by the end-users. Of the methods suggested by FEMTI to measure well-formedness, we use the method that identifies errors in the output, and categorises these errors according to Flanagan's (1997) classification. We also use a rating scale measuring the severity of the errors noticed by subjects.

The end-user evaluation presented in this study reuses elements of previously conducted human MT evaluation studies. We gather human judgements on texts (subtitles) using scales, building on work by Pierce et al. (1966), Van Slype (1979), Callison-Burch et al. (2007); we also administer a questionnaire to the subjects to gather opinions on the intelligibility and acceptability of machine-generated subtitles (JEIDA 1992). In addition to the human evaluation we evaluate the MT output with automatic metrics used extensively in current MT evaluation. The quality characteristics measured in the study and the metrics used to measure them were suggested by FEMTI based on the type of evaluation we are conducting. We already mentioned that the FEMTI model does not include recipient evaluations, and therefore we broadened our methodology to include a recipient evaluation based on the work of Trujillo (1999). This type of evaluation requires the recipients of the MT output to evaluate the translation in terms of quality, cost and speed. We adapt the methodology to focus on evaluating MT quality in terms of intelligibility and acceptability from the end-users' perspective. The questionnaire in the current study includes a mixture of open and closed questions, in addition to rating scales normally used in manual evaluation of MT. The mix of open and closed questions gathers data relevant to the four quality characteristics being examined. Even though we present a methodology for human evaluation of machine-

translated texts, the design of the evaluation does not follow that of the more 'traditional' human-based evaluations, and combining an interview with a questionnaire is a novel approach both in MT research and viewer-based subtitling studies. It allows us to gather sufficient amounts of data and to avoid many of the pitfalls associated with self-administered questionnaires and online surveys, a topic we discuss in more detail in Chapter 3.

## 2.3 Concluding Remarks

This chapter provided an overview of the general approaches to MT evaluation, highlighting the many influential studies that have shaped the way evaluation is conducted today. The chapter has shown there are many ways to construct an evaluation methodology, and that the context of use plays an important role in the evaluation when choosing quality characteristics and associated methods. A major choice point in an evaluation methodology involves deciding whether or not to conduct a human or automatic assessment of MT quality. We chose to conduct a large-scale human evaluation, which has not been conducted to date in the area of automated subtitling. We defined all the quality characteristics employed in our evaluation model, and the methods used to measure these characteristics. We introduced FEMTI, a framework being developed to standardise MT evaluation methodologies. This framework can be used to generate an evaluation strategy. It is a context-based evaluation tool, relating the quality model used to evaluate the machine translation system to the purpose and context of the system. However, the FEMTI model did not consider acceptability from end-users' perspective as a quality characteristic, given the text-based focus of FEMTI, and therefore it was not suitable in this study as a stand-alone model. We combined a declarative evaluation generated by the FEMTI model with a recipient evaluation (Trujillo 1999) asking end-users of the subtitles to evaluate them. We provided the rationale for our choice of evaluation methodology in this study, outlining a structured two-phase approach. An important aspect of our approach is taking the end-users of the subtitles into account. To date only a handful of MT evaluations have considered the end-users as an influencing factor when devising their quality characteristics and methods (JEIDA 1992, Armstrong et al. 2006c and Roturier 2006).

In the next chapter, we move on to discuss in more detail the research methodology employed in the current study. Chapter 3 introduces the theoretical and practical aspects

of the research design. Following this it outlines the corpus compilation for the study, and the EBMT system we used to generate the subtitles. It also provides details on both phases of our evaluation methodology, and discusses the setting of the end-user evaluation sessions. Lastly we present details of the data collection and analysis techniques used to measure the intelligibility and acceptability of EBMT-generated subtitles.

# - Chapter 3
# - Methodology

# 3 Methodology

The purpose of this chapter is to describe and explain the methodology used in this study. The chapter is organised in the following way: we look at the concept of research, the main and subsidiary research questions of this study, the overall design of the research, and we define two important terms when conducting any type of research: quantitative and qualitative (3.1). Following this we provide a detailed description of the EBMT system we used in the study (3.2) and we situate the system among other Corpus-Based MT systems. We then move on to introduce the principles of research design dealing firstly with conceptual issues including independent and dependent variables; operationalisation; units of analysis; internal, external and measurement validity; details of participants; questionnaire design and interview design (3.3). Following this we look at the practical issues involved in the research design (3.4). First of all we discuss why and how we compiled the corpora for use with the EBMT system, before discussing three important aspects of the study: subject selection, text selection and the design of the interview questionnaire. The final section (3.5) focuses on the practical implementation of the research, with particular focus on the two-phase approach taken in the current study.

## 3.1 Research in this Study

In this study we employ an individual interdisciplinary research model (Gebremedhin & Tweeten 1994:25), spanning the disciplines of computer science and translation studies. We employ a mixed methods research design,[55] the idea of which is to combine quantitative and qualitative approaches in many phases of the research process. Using quantitative and qualitative approaches in combination enables a better understanding of research problems than either approach alone (Creswell & Clark 2007). The research methods we employ under the umbrella of the mixed methods approach in this study are survey research, evaluation research and corpus-based research. When we discuss the various stages of our research methodology we will expand on these types of research.

[55] Over the past 50 years this name has been the topic of much discussion. Other names include integrated methods, combined methods, quantitative and qualitative methods, hybrids, methodological triangulation, combined research and mixed methodology (Creswell & Clark 2007).

According to Williams & Chesterman (2002:64) the aim of quantitative research is to be able to say something about the "generality of a given phenomenon or feature, about how typical or widespread it is, how much of it there is; about regularities, tendencies, frequencies, distributions." The aim of qualitative research is to describe the quality of something in an informative way. More importantly, "qualitative research can lead to conclusions about what is possible, what can happen, or what can happen at least sometimes. However, it does not allow conclusions about what is probable, general, or universal" (ibid). The data collected through qualitative methods will augment our understanding of what is possible in the particular research domain. Combining the two types of research can be used to achieve triangulation, in which different sources of evidence are used to shed light on research findings, and this practice increases the measurement validity and reliability of a study. 'Valid' measurements actually measure the attribute that they claim they measure. 'Reliable' measurements involve measuring the attribute in a consistent and stable manner. For a study to be valid it must also be reliable. For this study it is essential that valid and reliable measurement techniques are developed and employed when collecting and analysing the data. Relating this to our study of MT evaluation, we must ensure the corpus analysis techniques in the first phase (prospective) of our two-phase approach and the interview questionnaire design (used in first and second (retrospective) phases) are valid and reliable. In order to minimise errors in both phases, we systematically conduct the analysis on the corpus, focusing on repetition levels, and for the design of the interview questionnaire we build on the work of Armstrong et al. (2006c).[56]

### 3.1.1 Research Questions

We outlined the main research questions in the Introduction, but come back to them briefly here before we discuss the methodology in detail. As this is an interdisciplinary study there are two main research questions formulated in the following way:

[56] The work reported on in Armstrong et al. (2006c) describes the pilot study used for the current study.

**RQ1**: Are target language subtitles produced by an EBMT system considered intelligible and acceptable by viewers of movies on DVD?

**RQ2**: Is there a relationship between the 'profiles' of corpora used to train an EBMT system, on the one hand, and viewers' judgements on intelligibility and acceptability of the subtitles produced by the system, on the other?

We have mentioned previously that a pre-analysis stage is not normally conducted with the corpora used in Corpus-Based MT systems (cf. 2.2.1). The present investigation could be very beneficial to developers of Corpus-Based MT systems, as it could indicate whether the number of SL repetitions in the corpus plays a significant role in the acceptability of MT output, rather than simply the size of the corpus being the deciding factor, which is considered to be the case in most evaluations of Corpus-Based MT systems. The present study uses three corpora of increasing size, and containing an increasing number of SL repetitions, and of a varying degree of homogeneity and asks what effect these three factors have on the intelligibility and acceptability of DVD subtitles. We previously characterised a corpus as homogeneous on the basis that it contained only subtitles. This is, of course, a very limited view of 'homogeneity'. We would now like to view our subtitle corpora on a cline from homogeneous to heterogeneous based on the broad genres (fantasy, action, mixed) they represent, and the individual movie whose subtitles they contain. As we are re-subtitling movie clips from a *Harry Potter* movie (fantasy genre), introducing subtitles from different films, including *Lord of the Rings* (also fantasy) and a mix of movies from numerous other genres, we are weakening the homogeneity of the corpora, as we expand the size. When analysing the results we investigate the relationship between the size of the corpus, the number of SL repetitions, the homogeneity of the corpus, and the intelligibility and acceptability of the subtitles generated by each of the three corpora.

In addition to addressing RQ1 and RQ2 above during the course of the research we will draw some conclusions about three related subsidiary research questions:

**RQ3**: If the viewer understands the soundtrack, are they more accepting or less accepting of the EBMT subtitles?

**RQ4**: If the viewer has a 'linguistic background', are they more accepting or less accepting of the EBMT subtitles?

**RQ5:** If the viewer has prior knowledge of the movie or related material such as books, are they more accepting or less accepting of the EBMT subtitles?

### 3.1.2 Research Design

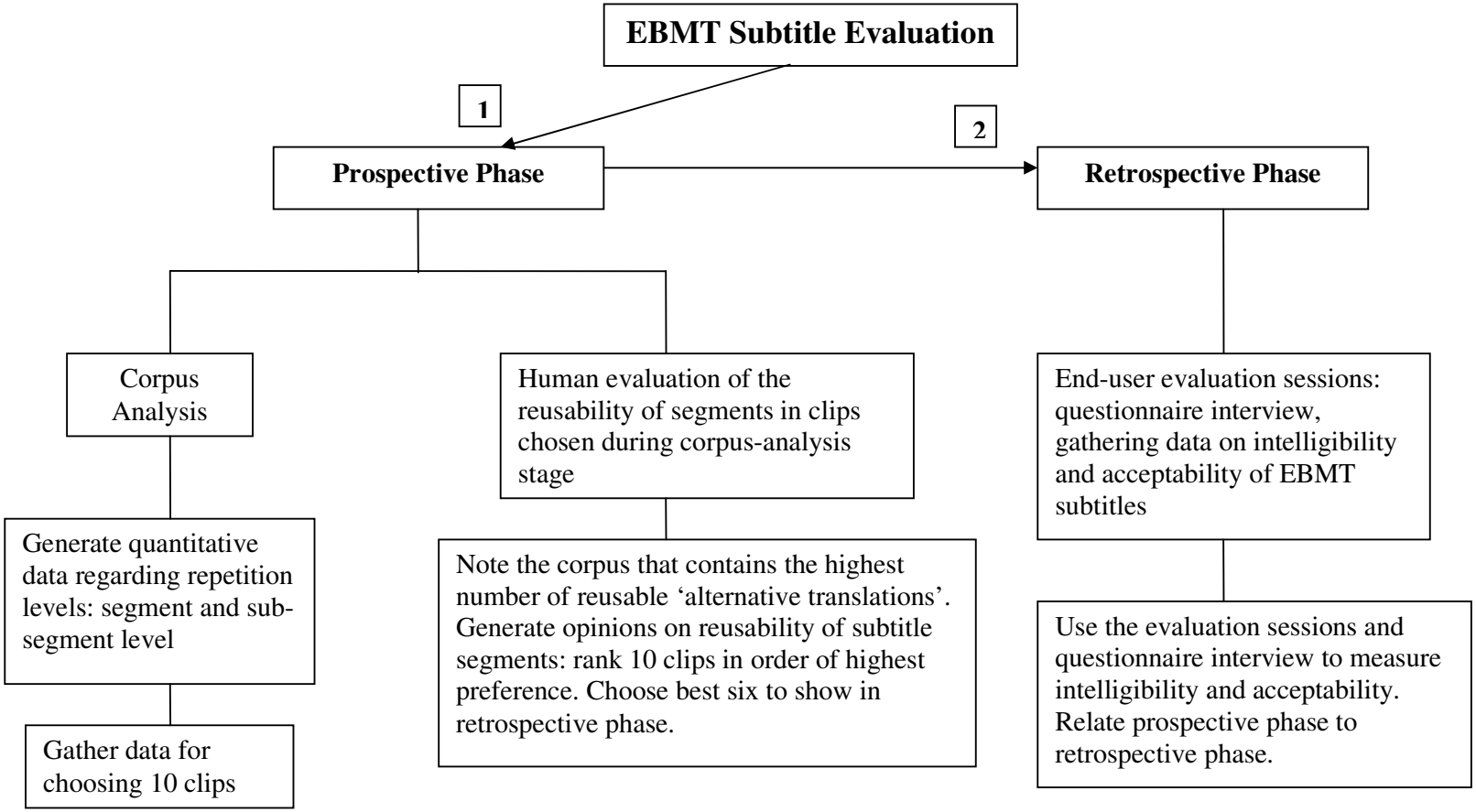
The research design (see Figure 3.1 for a complete overview) employed in this study is made up of a two-phase study, which builds on previous work conducted by Armstrong et al. (2006c). As indicated in Chapter 1, section 1.5.6, this earlier pilot study tested the feasibility of generating subtitles using an EBMT system on the one hand, and developed and tested two types of evaluation methods, namely face-to-face evaluation sessions and online surveys, on the other. It also tested specific software relating to corpus research (SDL Trados' Translator's Workbench) and the process of subtitling specific movie clips (DVD Decrypter, Avi2DVD and Subtitle Workshop).[57] Pilot studies provide an opportunity to check the stability of the technology and to minimise any flaws in the research design. The study used feasibility and usability evaluation methods previously mentioned in section 2.1.1. The study also employed survey research methods. Survey research uses methods that gather descriptive information about populations too large for every member to be studied and it is used in applied research to measure public opinion. We follow the notion of an interview questionnaire as outlined by Oppenheim (1992:102). Of the many types of interviews, the one that is most pertinent to this study is the research type interview. It consists of three interacting variables: the respondent (subject), the interviewer (researcher), and the interview schedule or questionnaire. As all of these variables including the interview location/situation have an influence on the results, it is important to control for as many as possible to ensure reliability and validity of the results.

[57] DVD Decrypter allows us to select the movie clips from the entire movie. AVI2DVD allows us to transfer the clips with subtitles onto a DVD. Subtitle Workshop is a freeware subtitling editing tool: http://www.urusoft.net/downloads.php?lang=1 [Accessed 10 March 2009].

An advantage of conducting an interview questionnaire rather than a standard individual self-administered questionnaire, a group-administered questionnaire or a mail questionnaire in this study is the researcher's ability to clarify any misunderstandings the subjects may encounter and thereby avoid losing any data due to subjects skipping questions, a problem that occurs most often with mail questionnaires and to a lesser extent with the other two types (ibid:33-34). Another advantage is the richness and spontaneity of information collected by an interviewer, something which the other questionnaires fail to obtain. (ibid:32). A possible disadvantage of conducting an interview questionnaire is bias introduced by the researcher, and this is discussed later in section 3.3.4. An online survey might have been a solution to remove some of the negative points associated with a researcher administering the interview questionnaire. However, following technical difficulties experienced by Armstrong et al. (2006c), coupled with the fact that an online questionnaire would not have provided the same volume of data, meant that we did not pursue this approach in the current study.

Building on the work of Armstrong et al. (ibid), we devised the current study, which is split into two phases: a small-scale prospective study and a larger-scale retrospective study (Figure 3.1 overleaf).

**Figure 3.1:** Overview of Methodology in this study



### Overview of Prospective Phase

Firstly, we will discuss the prospective phase, which is broken down into two stages:

- Corpus analysis
- Human evaluation of the reusability of TL subtitles

The corpus-analysis stage employs corpus-based research, an empirical research method that uses a corpus (or corpora) to analyse a particular phenomenon. A corpus is a large collection of naturally occurring texts (usually held in electronic form) and in the area of corpus linguistics, language is studied based on such corpora. In the case of a bilingually-aligned subtitle corpus, the corpus contains examples of actual subtitles created by professional subtitlers and the corresponding translations by the subtitler. As already indicated, nowadays corpora are usually found in electronic or machine-readable format. For this reason the discipline of corpus linguistics makes use of computer technology and corpus-processing software, making the process of data manipulation easier than if the data was processed by hand. There is extensive corpus-based research conducted in the disciplines of translation studies and corpus linguistics (Baker 1995, Biber et al. 1998, Kennedy 1998, McEnery & Wilson 2001, Kenny 2001, Olohan 2004). In contrast to the type of studies previously conducted, corpus-based research in the current study is not intended to investigate a particular linguistic phenomenon. The purpose of the corpus analysis phase is to locate source language (SL) repetitions in the corpora and their corresponding translations in the target language (TL) to investigate RQ2 outlined above and to collect the necessary data we require for the retrospective phase. During this phase we still follow standard procedures of corpus-based research by using software to investigate the corpora in order to manipulate data and generate statistics for the repetitions. As outlined earlier Volk & Harder (2007), Volk (2008) and Hardmeier & Volk (2009) conduct corpus profiling which provides details on the characteristics of the corpus used in their studies. Building on this work we provide similar details in addition to information on our independent variables and we describe the practical implementation of the corpus-analysis stage in section 4.1.[58]

[58] As regards corpus creation and corpus analysis, dialogue lists (cf. section 1.2.3) could be useful in future studies. It is now possible to download the lists in electronic format, allowing researchers to easily investigate the content using corpus-analysis techniques. Dialogue lists, for example, could give us information on source text subtitle repetition and subtitle length (in cases where the 'master' subtitles have already been provided). They also allow us to investigate pre - and post-production movie dialogue.

The second stage in the prospective phase of our study, human evaluation of the reusability of TL subtitles, employed survey and evaluation research methods to gather subjective opinions on the reusability of subtitles using an interview questionnaire. This stage is a follow-up to the corpus-analysis stage, as the subjects made judgements on the subtitles of movie clips chosen in the corpus-analysis stage, and based on this data we chose our movie clips for the retrospective phase. The practical implementation of this second stage is described in section 4.2.

### Overview of Retrospective Phase

Secondly, we discuss the retrospective phase, which consists of end-user evaluation sessions. Once again this phase uses survey and evaluation research methods as outlined above, including the use of interview questionnaires in a face-to-face environment, to collect data. The evaluation model used in the retrospective phase is based on the work of FEMTI and a recipient evaluation as outlined in the previous chapter. The practical implementation of this phase is outlined in sections 5.1 and 5.2.

Before we move on to discuss the specific theoretical and practical aspects of the study, we will give a more in-depth description of the EBMT system mentioned in the previous chapters, and which we use in this study to generate the subtitles.

## 3.2 The EBMT System

The system used in this study is called MaTrEx (Machine Translation by Example). It is essentially a hybrid 'example-based SMT' system (Groves 2007), originally developed in the National Centre for Language Technology (NCLT) at DCU in early 2006. Since then new methods have been added to improve system quality. We used the system in June 2007 to generate the output which is used in the end-user evaluation sessions, and therefore will only describe the system as it was at this time.[59] The system is a modular data-driven MT engine, built following established design patterns (Gamma et al. 1995) and consisting of a number of extendible and re-implementable modules. The four most important of these modules are outlined by Stroppa & Way (2006):

[59] For information on updated versions of the system, see Hassan et al. (2007), Tinsley et al. (2008), Du et al. (2009). This EBMT system is also being used as a background IP in the Centre for Next Generation Localisation, which is a Centre for Science Engineering and Technology based at DCU. There was a major redesign of this system which began in November 2008, leading to a number of improvements being integrated.

*Word alignment module*: This module takes as its input a segment-aligned corpus and outputs a set of word alignments

*Chunking module*: This module takes as its input a segment-aligned corpus and produces source and target chunks

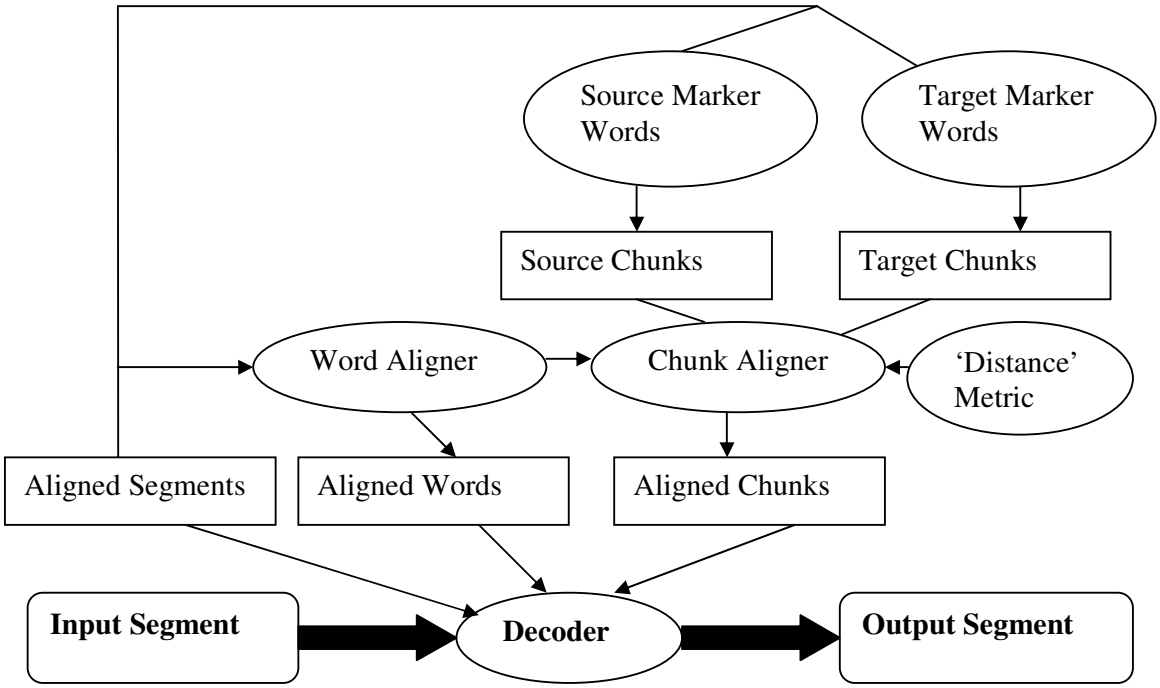
*Chunk alignment module*: This module takes in the source and target chunks and aligns them on a segment-by-segment level

*Decoder*: This module searches for a translation to a new input using the original aligned corpus and derived chunk and word alignments

In Chapter 1 we gave a cursory description of EBMT. Here we elaborate on some of the concepts previously mentioned and outline further aspects of the system.[60] Figure 3.2 gives the reader a simplistic overview of how the system functions: the aligned source-target segments are passed in turn to the word alignment, chunking, and chunk alignment modules, in order to create the chunk and lexical example databases. The three databases are then given to the decoder to translate new input segments (ibid:32). An important thing to observe here is the modular design implemented in the system. This means that all the main modules mentioned above can be easily re-implemented, extended or adapted to allow the integration of existing software, such as the use of wrapper technologies, a technique whereby an interface is created around an existing piece of software (Groves 2007:113).

[60] An in-depth analysis of the MaTrEx system is outside of the scope of this thesis, and therefore the reader is advised to consult the previously published literature on the MaTrEx system mentioned above.

**Figure 3.2:** The MaTrEx Translation Process




At this point we define what a segment is, as this is the unit of analysis used in the current study when considering the acceptability of EBMT subtitles. We describe the corpora as being sententially aligned. This means that each subtitle in English is aligned with the corresponding translated subtitle in German. A subtitle could be described as a segment rather than a sentence, as subtitles can range in length from one word to fifteen words. A sentence is thought of as a string of words beginning with a capital letter and ending with a punctuation mark. A segment, however, can include sentences as well as what we might call utterances, for example, 'No way!' or 'Stop that!' which are commonly occurring subtitles. Therefore the corpora are aligned at segment-level based on sentential punctuation.

### 3.2.1 Chunking

We have mentioned the word 'chunking' and 'chunks' on numerous occasions when referring to the EBMT approach to automatic translation. A chunk is a segment which is

part of a larger segment that has been broken up by the algorithms used in our system. There are different approaches to EBMT in respect to how the segments are broken up into sub-segments or chunks in order to find matches in the aligned corpus. The approach taken with the MaTrEx system is to use the Marker Hypothesis, which states that:

> All natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context (Green 1979:483).

The idea behind this is that languages are marked for syntactic structure at the surface level and can therefore be broken up into smaller, yet still useful segments or chunks, which can then be recombined to form new correct sentences or segments. The marker words included in our system are: determiners <DET>, quantifiers <QUANT>, prepositions <PREP>, conjunctions <CONJ>, WH-adverbs <WH>, possessive pronouns <POSS-PRO>, personal pronouns <PERS-PRO>, and punctuation marks <PUNC>. Marker words are required for both languages, in this case English and German. Lists of marker words are first extracted from CELEX (Centre for Lexical Information) and then edited manually to ensure all categories have been included. It is also the case with German that we removed the reflexive pronoun as a marker word because when German segments were chunked based on this tag we found that the corresponding English chunk was not an equivalent translation, given that English does not contain the same reflexive pronoun (Example 3.1, subtitle taken from the movie *Frantic* (1988)):

**Example 3.1:** Reflexive Pronoun in German

| | |
|---|---|
| English: | <PERS-PRO>I brush <POSS-PRO> my teeth |
| German: | <PERS-PRO> Ich putze <REFLEX-PRO> mir <POSS-PRO> die Zähne |

In this (invented) example ‘my teeth’ would be aligned with ‘mir die Zähne’, which is a dative reflexive personal pronoun + the teeth. If this alignment is saved in the training corpus, each time ‘my teeth’ is presented as input, an incorrect translation would be offered. There are many examples of this kind of reflexive verb in German, so, on the advice of the current researcher, the EBMT developer agreed to keep the reflexive

personal pronoun with the verb chunk, for example 'Ich putze mir' is aligned with 'I brush'.

The Marker Hypothesis is first applied in a pre-processing step where the source-target segment pairs are tagged automatically with their marker categories. The aligned source-target chunks are then generated by breaking up the segment based on the tags, as well as using word translation probability and cognate information. Each chunk must contain at least one non-marker word to ensure the chunk contains useful contextual information (Gough & Way 2004). The following example of a subtitle from *Harry Potter and the Philosopher's Stone* (2001) shows that 'in' would not be considered a chunk on its own, as it is followed by another marker word 'the'. Instead, the two marker words are joined together and the chunk would then contain two marker words and one non-marker word (Example 3.2):

**Example 3.2:** A chunk must contain at least one non-marker word

| | | |
|---|---|---|
| Troll <PREP>in <DET>the dungeon | → | <CHUNK>in the dungeon |

### 3.2.2 The MaTrEx System – Hybrid Approach

As mentioned previously in Chapter 1, corpus-based or 'data-driven' approaches to MT dominate the MT research agenda. EBMT and SMT are the two different frameworks within the data-driven paradigm, and of these SMT is the more dominant type with many readily available SMT systems including free online MT systems, such as Google Translate.[61] On the other hand EBMT systems are relatively unavailable within the commercial sector.[62]

We also noted that early SMT models used only word-level correspondences (Brown et al. 1990, 1993), but SMT has since progressed to using sophisticated phrase-based approaches (Koehn et al. 2003), whereas EBMT approaches to MT have always made use of both phrasal and lexical correspondences (Nagao 1984). This has caused the boundaries of the two frameworks to become slightly blurred, meaning it is now more difficult to distinguish between them. It has been recognised that both approaches have

[61] http://www.google.com/language_tools?hl=en [Accessed 10 March 2009].
[62] We already mentioned one EBMT system, Traslán which is being used commercially.

strong and weak points, depending on the training data and domain, and therefore combining the two could possibly increase the quality of the output. Hutchins (2005) notes that given that EBMT uses both statistical (SMT-like) and linguistic (RBMT-like) methods, it is difficult to characterise and define EBMT, and suggests that perhaps EBMT is already a true 'hybrid' approach. In agreement with Hutchins, Groves & Way (2006a) comment that combining both SMT 'phrases' and EBMT 'chunks' in either a 'statistical EBMT' or 'example-based SMT' system will improve the system output, and adhering solely to one method will hinder the performance of the system. Groves & Way (ibid) have conducted experiments using automatic metrics to test the influence a combination of the different approaches has on the automatic scores, and the results show a clear improvement when both statistical and linguistic methods are used together.

Groves's (2007) work on hybrid data-driven systems gives an excellent description of how the different frameworks perform the steps involved in MT. He conducts translation experiments to compare the different system combinations, generating interesting results concerning the use of hybrid data-driven approaches. His results suggest EBMT is particularly suited to translation within a sublanguage domain such as that represented by *Sun Microsystems* data; however, SMT is more suited to translation from a more open domain such as the Europarl corpus (Koehn 2005). Introducing EBMT chunk alignments into the SMT system resulted in improvements over the baseline system, indicating the higher quality of the EBMT chunk alignments over the SMT phrasal alignments. The overall results suggest that the crucial differences between the two approaches contribute positively to the overall translation quality (ibid:146).

As already adverted to in Chapter 1, one of the most significant advantages EBMT systems have over their SMT system counterparts is the ability to store examples in the training corpus, thereby allowing these examples to be re-used during translation and to be considered exact matches in subsequent uses of the system (see Way & Gough 2005a). EBMT systems do not carry out the chunking and recombination stages all over again. In contrast to this when an SMT system encounters a previously translated segment during the translation stage it repeats the same steps to find a translation which maximises the probabilities of the translation and language model, thus increasing costs

in relation to time and effort. Groves (ibid:48) notes that despite exact sentence matching being a relatively simple exercise, it is not evident in the literature published to date that this technique is in fact being used in any of the current SMT approaches. The MaTrEx system implements this exact sentence matching process, allowing a TL translation to be retrieved in its entirety and by-passing the recombination step.

#### 3.2.2.1 Locating MaTrEx among other Systems

In Chapter 2 we described MT shared tasks, the aim of which is to evaluate MT systems using a pre-defined data set. There has generally been a lack of participation among EBMT researchers in competitive evaluations, meaning there is little research comparing EBMT and SMT systems available. However, this trend is changing with the MaTrEx system having competed in numerous tasks to date including: OpenLab 2006, IWSLT-06, IWSLT-07, IWSLT-08, ACL-08, ICON-08, ACL-09, finishing competitively in each one (see Armstrong et al. 2006a, Stroppa & Way 2006, Hassan et al. 2007, Ma et al. 2008, Tinsley et al. 2008, Srivastava et al. 2008, and Du et al. 2009 for further details on the system and its performance).

Describing the MaTrEx EBMT system as a hybrid data-driven MT system means that it uses both EBMT (linguistic) and SMT (statistical) approaches to extract aligned chunk resources. Below we briefly outline the different strategies employed within each module.

#### 3.2.2.2 Alignment Strategies

*Word Alignment Module*: This uses the statistical word alignment tool GIZA++[63] to extract word and morpheme alignments, which are then passed to the translation decoder.

*Chunk Alignment Module*: In order to align the chunks extracted during the chunking phase based on the Marker Hypothesis, an edit-distance style alignment algorithm is implemented, which computes the most likely alignment between two chunks (Stroppa & Way 2006, Stroppa et al. 2006).

[63] <http://www.fjoch.com/GIZA++.html> [Accessed 10 March 2009].

*Integrating SMT Data into Chunk Alignments*: SMT data are integrated into the EBMT system by adding SMT phrasal alignments to the aligned chunks extracted by the chunk alignment module, in order to improve the quality of the output translations (ibid), as recent research (Groves & Way 2005, 2006a, 2006b) has shown that systems that combine the alignment techniques of both data-driven approaches have outperformed the baseline systems from which they are derived.

#### 3.2.2.3 Decoder

The decoder is also a hybrid system, integrating EBMT and SMT. It is capable of retrieving already translated segments, while also providing a wrapper around PHRAMER,[64] a phrase-based SMT decoder.

#### 3.2.2.4 Pre-Processing Stage

Before we train or test the system there is a pre-processing stage during which we prepare the data. Both the training corpora and the test data have to be processed in the same manner so that all the data are in the same format. The text is prepared by tokenising (e.g. separating punctuation symbols, including hyphens, from words) and by converting all uppercase letters to lowercase. The pre-processing stage is conducted to make the data compatible with the system algorithms. Separating punctuation symbols and words makes it easier for the system to recognise words and using text that is all lowercase allows the system to recognise, for example, that the word 'The' is the same as the word 'the' during the translation process.

### 3.2.3 Training the System

The system is trained on three separate bilingually-aligned corpora (English-German) to produce three sets of output, as outlined below in section 3.4.1. The EBMT system is trained on each of the corpora, and training is completed in very short amounts of time ranging from one minute for the first corpus to a maximum of three minutes for the third corpus.[65] Similarly the other steps within the pre-processing stage are completed in short time-frames. The number of subtitle pairs we use in our training data is very low

[64] <http://smt.phramer.googlepages.com/> [Accessed 10 March 2009].

[65] The MaTrEx system is easily accessible via a server within DCU, which facilitated the current research. We had full access to the system in terms of selecting/deleting the training corpora and test data, and we had the option of generating automatic metric scores for the MT output. We did not, however, have access to the internal workings of the EBMT system. That said, the system developers were always available for questions and they were extremely helpful.

compared to the training data used by Volk & Harder (2007) and Volk (2008), who use 4 million subtitle pairs in their training data set in comparison to our three data sets of 6,997 11,342, and 42,331 subtitle pairs respectively. This is due to the fact Volk & Harder (ibid) carry out their research in conjunction with a Swedish subtitling company, and are thus able to avail themselves of such large data sets. In addition, Volk & Harder (ibid) use subtitles for TV programmes, including soap operas, detective series, documentaries, and feature films as their training and test data (ibid:452), which is a larger domain from which to extract subtitles.[66]

Now that we have described the EBMT system we used to generate the subtitles for the retrospective phase of this study, we turn our attention to the theoretical aspects of the methodology that needed to be considered for this study.

## 3.3 Theoretical Research Design

Every research design contains a theoretical and a practical element. In the theoretical element there are basic concepts we need to consider before we can implement a practical research design. In section 3.3.1 below we draw on the principles of research design developed by Frey et al. (1991), Bailey (1994), Oppenheim (1992) and Williams & Chesterman (2002) and examine in particular variable relationships, operationalisation, unit of analysis and internal, external and measurement validity. In section 3.3.2 we outline our practical research design, including the design of the interview questionnaire, our primary "retrospective" method of data collection and analysis.

### 3.3.1 Variable Relationships

In any kind of research there are various relationships that may exist between variables, either before the study is conducted or perhaps only on completion of the study. Bailey (ibid:47) discusses symmetrical and asymmetrical relationships: *symmetrical relationships* mean that a change in either variable is accompanied by a change in the other variable. In contrast, with *asymmetrical relationships*, a change in variable A is accompanied by a change in variable B, but not vice versa. Following on from this, in

[66] During the pilot study phase of the research (Armstrong et al. 2006c), we discussed with a subtitling company who subtitle mainstream Hollywood films the possibility of obtaining large amounts of subtitles from them, but unfortunately subtitling companies normally have to transfer the copyright of the subtitles to the distributor and therefore do not own it themselves (Mary Carroll; personal communication).

an asymmetrical relationship, the variable that is capable of effecting change in the other variable is called the *independent variable* and the variable whose value is dependent upon the other but which cannot itself affect the other is called the *dependent variable*. In the present study we can identify the independent variables as the number of SL repetitions in the bilingually-aligned corpora, the size of the corpora and the level of homogeneity in the corpus; and the dependent variables as the intelligibility and the acceptability of EBMT-generated subtitles. We want to investigate whether or not an increase in the number of SL repetitions, the size of the corpus and the homogeneity of the corpus increases the intelligibility and acceptability of the subtitles from the point of view of the end-users. This will be of particular interest to the MT community when gathering data for corpus compilation. Bailey (ibid) and Frey et al. (ibid) highlight that a causal relationship can be established between two variables if they adhere to three particular requirements. Therefore, X causes Y if:

- There is a relationship between X and Y.
- The relationship is asymmetrical (X precedes Y).
- The changes observed in Y must be the result of changes in X and not some other, unknown variable. (This ensures that the causal relationship is a valid one, rather than the result of various internal validity threats outlined in section 3.3.4.1.)

### 3.3.2 Operationalisation

According to Frey et al. (ibid:94 and see Williams & Chesterman 2002:78) every research study must define the observable characteristics of the concept they are interested in, as it is not possible to measure the abstract concept directly. This process is referred to as *operationalisation* and the process must involve three characteristics: firstly, the definition must be adequate, providing a complete description of the characteristics being observed; secondly, the definition must be accurate, meaning it is valid and universally agreed; and thirdly, the definition must be clear to the readers and to future researchers.

In this study we want to measure the intelligibility and acceptability of automatically-generated subtitles, but we are unable to measure these characteristics directly. We therefore specify observable characteristics of intelligibility and acceptability. We noted

in the previous chapter that research on intelligibility and acceptability of MT output includes Carroll (1966), Dostert (1973), Van Slype (1979), Halliday (cited in Van Slype 1979), Trujillo (1999), Popowich et al. (2000) and Bowker & Ehgoetz (2007). Our approach is to elicit subjective judgements on whether subtitles exhibit the following quality characteristics, which we have previously defined: comprehensibility and readability (both contributing to intelligibility), and style and well-formedness (both contributing to acceptability) and we measure these characteristics through the use of qualitative and quantitative methods.

### 3.3.3 Unit of Analysis

The unit of analysis in this study is the subtitle. However, we do not analyse it in exactly the same way during each phase, and outline the differences below.

#### Prospective Phase

During the corpus-analysis stage (see section 3.5.1) the unit of analysis is both the segment (subtitle) and the sub-segment[67] which results from breaking down the segment into chunks of 2-4 words. Gathering statistics on sub-segment repetitions is also of interest in the overall study of the acceptability of the subtitles. The EBMT process relies a great deal on the breaking down of segments into sub-segments, and therefore gathering statistics on sub-segment repetitions in the movie clips in advance of run-time could possibly indicate to the researcher how useful each corpus might be. When gathering human judgements on the reusability of TL subtitles, we ask the evaluators to comment on each subtitle individually.

#### Retrospective Phase

During the end-user evaluation sessions we present the subjects with six movie clips, with each clip containing between 23 and 36 subtitles. The unit of analysis in this phase is the group of subtitles on the movie clip, and not an analysis of individual subtitles. We want to provide as natural a setting as possible and asking subjects to give judgements on a group of subtitles is more appropriate in an AVT context.

[67] We are unable to use the entire corpus for locating sub-segment units, as the data set is simply too large and a manual study of the entire corpus would not be feasible. Therefore we look at sub-segments in the ten movie clips chosen for the next stage only.

### 3.3.4 Internal, External and Measurement Validity

Frey et al. (1991) identify two types of validity: *internal validity* and *external validity*. In addition to these two types Frey et al. also mention measurement validity and its associated concept of measurement reliability, all of which we will discuss in this section.

#### 3.3.4.1 Internal Validity

Internal validity concerns the accuracy of the conclusions drawn from a study. There are many factors which can threaten the internal validity of a study and Frey et al. identify three particular categories: threats due to researchers, threats due to how research is conducted, and threats due to research subjects (ibid:125). Some of these categories identified by Frey et al. apply only to methods involving subjects. If the category is not applicable to the methods involved in analysing the corpus, then we do not include references to this category in what follows.

##### Threats due to Researchers

The first threat concerns the influence a researcher may have on the participants in the evaluation. This is known as the *researcher effect* (ibid). The researcher effect can be split into two categories: the *researcher personal attribute effect* and the *researcher unintentional expectancy effect*.

*Researcher Personal Attribute Effect*

This effect occurs in a study if the characteristics of a researcher influence peoples' behaviour or answers to a questionnaire, for example. Frey et al. report on previous research (Barber 1976, Yagoda & Wolfson 1964) that has shown how particular characteristics of a researcher can influence a subject's response (e.g. gender, age, race, friendliness). The research noted that this effect is likely to occur under two conditions: when the research task is ambiguous the subjects tend to look to the researcher to indicate how they should perform, and when the task is related to the personal characteristics of the researcher, for example, if the researcher has particular religious beliefs, the subject may feel they need to respond in accordance with these beliefs.

*Researcher Unintentional Expectancy Effect*

This effect occurs in a study when researchers influence a subject's response by unconsciously indicating to the subject the results they are looking for. Frey et al. report on research that looks at this effect (Rosenthal 1966), and an example of swaying the subject in one particular direction is a researcher smiling when they hear the answer they are looking for and frowning if they don't.

In both cases Frey et al. mention two possible ways of controlling for the effects. The first suggestion is to employ a wide variety of research assistants, removing the researcher who is conducting the overall study, so that they cannot influence any of the subjects' responses. Another advantage of employing numerous research assistants is that assistants can be matched up with subjects they are least likely to influence. However, this cannot guarantee that a subject will not be influenced by the assistant. The second suggestion is to follow a standard procedure therefore ensuring that each subject is exposed to the same research environment.

In order to control these effects in this study, it was not feasible to employ numerous research assistants. Therefore we opted for the second suggestion and followed a standard procedure for each evaluation session, including the two evaluation phases in the prospective phase and the evaluation sessions within the retrospective phase. In all instances we kept as many factors as possible constant including the research environment, pre-viewing briefings, the language of the subtitles, the format of the sessions, and the length of the sessions. The pre-viewing briefing outlined the background to the research, the format of the session, and instructions the subjects should follow. Using a pre-written briefing during each session meant subjects received identical information and clear instructions. The briefing was written in such a way that the researcher's preferred or expected results were not communicated to the subjects. If this did happen, then it was unintentional.

## Threats due to how research is conducted

The second threat identified by Frey et al. (ibid) is one that is introduced by how the research is conducted. Of the factors influencing this threat, the *validity and reliability of the procedures used*, *sensitisation* (see White (2003), who uses the term 'testing effect' as a synonym for sensitisation), and *data analysis* (ibid:126) are relevant to the current study.

*Validity and Reliability of the Procedures used*

The first of these factors, the validity and reliability of the procedures used, requires the researcher to apply the research procedure consistently and accurately. In the cases of interaction with subjects, this is achieved by ensuring subjects are exposed to the same levels of variables and by using accurate measurement techniques in a consistent manner. In this study all subjects were provided with the same pre-viewing briefing, the evaluation sessions were situated in the same location, each subject was asked to carry out the same task, and the same measurement techniques were used for all the subjects.

In the corpus analysis phase (prospective phase) we used software to analyse the corpus and to generate statistics. Here accurate and consistent measurements were ensured by careful monitoring of the system settings (e.g., making sure that the same segment boundaries were consistently applied in SDL Trados' Analyse tool) and of the data currently under investigation (through file and translation memory selection). In both the prospective and the retrospective phases we used interview questionnaires. Using interview questionnaires posed the possibility of bias being introduced due to question-wording. To combat this when we designed the interview questionnaire we introduced factual questions and attitude questions, making it necessary to establish the reliability and validity of both question types.

**Factual questions** usually have a true answer, so asking the same factual questions over and over again in similar situations should yield consistent results, showing high reliability. Oppenheim (1992:145) maintains that one way of ascertaining the reliability of factual questions is to include internal checks in the interview questionnaire. Internal checks highlight any inconsistencies in the answers given by the subjects and they notify the researcher of any sources of error, for example faults in the wording of the questions. Example 3.3 below outlines internal checks used in one of the questionnaires from the main study. In this example we wanted to determine subjects' knowledge of the *Harry Potter* series before they viewed any of the movie clips. Asking them a closed question about if they have read any of the books or if they have watched any of the movies, followed by an open question asking how many of the books or movies they have read or watched, checks the reliability of their answer to the first question. The final question in this example then checks whether they would still be able to easily

recall information on the different characters from the books and/or movies since reading or watching them.

**Example 3.3:** Using internal checks to ensure reliability and validity

1. Have you read any of the Harry Potter books?
    a. Yes
    b. No
2. How many?
3. Have you seen any of the Harry Potter movies?
    a. Yes
    b. No
4. How many?
5. Do you know the characters in the books/movies?
    a. Yes
    b. No
    c. Not really – just the names, but not who they are

According to Oppenheim (ibid:146) external checks can also be carried out, whereby a second, independent source of information is required to validate the first source. However, it is also noted that this can be quite a difficult technique to employ and it is not a suitable solution in the current study. Another technique is the introduction of 'quality checks' into the study. This means that some of the respondents who have been interviewed in the usual way are later re-interviewed by a different group of trained interviewers. This is a suitable technique in situations where a large number of temporary interviewers are employed to help with the study, but this technique did not apply to this study (only one interviewer was involved in the evaluation). According to Oppenheim (ibid:90) a common mechanism to maximise the validity of a response is to establish a good rapport with the subjects, so that the subjects are more willing and eager to provide accurate information. This technique applied to the current study. When the evaluation sessions involved just the subject and the researcher, the subject might have felt slightly nervous at the beginning of the session. Therefore it was important that the researcher made the subject feel at ease in order to gather the most

accurate information. The current study employed the method of conducting a questionnaire in tandem with an interview as a technique to ensure reasonable validity.

**Attitude questions** differ from factual questions in that they are more sensitive to changes in wording, context and emphasis (Oppenheim 1992:147). Attitudinal questions cannot be asked in another form to assess reliability, as then it would no longer be the same question. Oppenheim suggests using sets of questions or attitude scales to avoid relying on single questions to measure the attitudes which are most important to the study. Sets of questions are considered more reliable than single questions as the sets of questions give more consistent results, as any bias caused by particular wordings can be cancelled out.

Assessing the validity of attitude questions poses some difficulties. Oppenheim (ibid:148) identifies some techniques which may help with this process. If in advance of an evaluation session we knew the attitude characteristics of the subjects, we might be able to predict certain behaviour, but this is not always the case. There is a complex relationship between attitudes and behaviour, often making it difficult to infer one from the other. Another approach to establishing the validity of attitude questions is to compare the findings of one study with the results of other studies in the same area. Such a technique could indicate that a study is on the right track if the various sets of results corroborate each other. On the other hand if the results are drastically different how does one establish whether one set of results is more valid than the rest? Comparing results with another study was not a viable option in the case of this study as there were no other end-user evaluation studies of automatically-generated subtitles in the literature.[68] The final technique identified stresses the value of the openness, depth, and intensity of the information obtained. It encourages subjects to respond to questions in an open manner, allowing them to take their time and state their views in their own way. This approach allows the researcher to obtain an overall picture of the subject's attitude to the issue at hand. The open question allows the subjects to express their attitudes more clearly than if they were given a choice of three different answers for example, which makes their answers more valid. The problem of validating attitude-

[68] The only comparison we can make at present is to compare the performance of the MT systems used to generate the subtitles, based on automatic metric scores, for example, those provided by Volk (2008) and Volk & Harder (2007).

type questions still remains a very real one (Oppenheim ibid:149), but one which we tried to overcome in this study by using open and closed questions uniformly.

### *Sensitisation*

The third factor, sensitisation or testing effect, refers to the fact that evaluators (subjects) often react differently to something the second time they see it than they had the first time: the initial measurement influences a subsequent measurement. White (2003:219) points out in relation to evaluating MT output that subjects who experience a badly translated expression may judge the next expression better than it really is, simply because the second translation was 'better' than the first in their opinion. For the study at hand sensitisation could have been considered a threat to the validity of the data collated. To counter this threat we organised the evaluation session so that the subject viewed a movie clip for two minutes, and then answered a set of questions relating to this clip.[69] The same format followed until all clips had been viewed. This way we gave the subjects a short break between each clip and by doing so reduced the possible influence of one clip on the next. We also provided short movie clips to minimise the effort required to recall the previously seen subtitles. During the retrospective phase evaluation sessions we showed subjects six movie clips, we generated three different sets of subtitles using three different corpora, and forty-four subjects participated in the evaluation sessions. When we were designing the evaluation sessions, we decided to alternate the soundtrack on the clips between English and Dutch,[70] to see if and to what extent knowledge of the soundtrack might have influenced subjects' judgements of the acceptability of the subtitles. Therefore there were six different sets of movie clips and subtitles. In all cases, the movie clips were shown in a chronological order as they appeared in the original movie, providing the subjects with some kind of storyline cohesion and reducing any misunderstandings due to the 'out of context' nature of viewing movie clips lasting only two minutes. This meant that between the forty-four subjects, fifteen subjects viewed subtitles generated by the EBMT system which was trained on Corpus AM, fifteen subjects viewed subtitles generated by the EBMT system

[69] This control was not implemented during the evaluation sessions in the pilot study, and therefore we observed the threat first hand (see Chapter 1). The control of viewing a clip and then asking questions was successfully implemented during the current study evaluation sessions.

[70] We would have preferred to choose a second language that is very different to German, but the *Harry Potter* movies available for purchase in region two only supplied English, German and Dutch as the possible soundtrack options. We recognise that the choice of language might have had an influence on the subjects' answers, as German and Dutch are related linguistically, but this is something that would need to be retested in the future.

which was trained on Corpus BM subtitles and fourteen subjects viewed subtitles generated by the EBMT system which was trained on Corpus CM (see section 4.1 for an explanation of 'AM', 'BM' and 'CM'). The set of movie clips and subtitles a subject watched was chosen at random and this reduced any bias on the part of the researcher and subjects. The subjects were not advised that there were different sets of subtitles or that the soundtrack was alternated depending on the particular set of movie clips. Unfortunately, it was not possible in this context to fully control the influence each clip had on the subsequent clips and we acknowledge that this could have introduced some bias.

#### *Data Analysis*

The last factor considered a threat to the internal validity of a study is identified as data analysis. When a researcher uses improper procedures to analyse data, it leads to invalid conclusions. In order to minimise the risk of invalid analysis, procedures must be conducted in a systematic way and the results recorded appropriately, e.g. in spreadsheets. We followed this guideline for our collection of the data during the corpus-analysis stage and for each of the evaluation sessions. We thereby ensured the risk of threatening the internal validity of the study remained at a minimum. We describe the data analysis procedures in more detail in the next chapter.

### Threats due to Research Subjects

In a research study subjects often pose a threat to internal validity. Of the six threats identified by Frey et al. (ibid), the following three are relevant to this study: *selection of subjects*, *maturation*, and *intersubject bias*. White (2003) also mentions maturation as one of the classic threats that can influence an MT evaluation study.

#### *Selection of Subjects*

When choosing subjects for a study, it is important to consider the various attributes they can introduce into a study and the effect they can have on the final results. The only prerequisites when recruiting subjects for this study were that all subjects evaluating the German subtitles had to be German native speakers and had to have seen at least one subtitled movie on DVD in the past.[71] Native speakers of a language have

[71] That said, however, the sessions were conducted through English, meaning if a person wanted to participate in the research study they would have to have an advanced level of English to answer questions during the interview questionnaire. All non-native English speaking students studying at an

the ability to easily recognise linguistic and structural errors in that language and they were the most suitable for the study at hand. In addition for each evaluation session we chose subjects of similar educational background (level of education), age group and knowledge of the English language. Within the groups there was a mix of gender and a mix of how often the subjects watched subtitled movies.

When we advertised for subjects, we specifically did not state that the subtitles to be evaluated were generated by an MT system. We did not choose subjects by whether or not they had any negative opinions on the subject of MT, as the results would not be representative of end-users of subtitles if we only recruited subjects who thought MT is of great benefit to subtitlers in advance of the sessions. We considered that the threat due to subject selection was reduced to a minimum after choosing subjects based on the criteria mentioned above. Later in section 3.4.3 we outline in more detail how we recruited subjects.

#### *Maturation*

Maturation refers to internal changes that occur within people during the course of an evaluation. Very ordinary things can affect someone's ability to be consistent in their judgements. As outlined by White (2003:219), subjects can get tired, bored, hungry or fed-up with the evaluation process, which in our case could mean subtitles evaluated later on in the session are evaluated differently to the subtitles graded at the beginning of the session. We minimised the risks mentioned above by keeping the sessions to a short length of time, and the subjects were aware in advance of their own time slots, allowing them to organise their day appropriately. We have also mentioned that the clips were shown in chronological order, therefore minimising any confusion the subjects may experience if they were not already familiar with the storyline.

#### *Intersubject Bias*

The final threat due to research subjects is intersubject bias that results when the subjects being studied influence one another. This was a clear threat to the study at hand. To counter this threat we briefed all subjects individually and conducted all the

---

Irish university must present a particular level of English from a recognised English language exam. Therefore we are confident of the minimum language level of the students. Even though the researcher has a good knowledge of German (spoken and written), she decided to conduct the interviews through English, as she did not consider her German advanced enough to conduct the interviews through German. Conducting the interviews through English also meant that there were no misunderstandings on the researcher's part and the data could be recorded more efficiently.

evaluation sessions on an individual basis. Subjects were asked to arrive at the evaluation venue at a specific time. The time-slots allocated were longer than the time the subject would require to complete the session, therefore reducing the risk of subjects meeting each other either before or after their individual session. Some subjects were studying the same course at university, and this could pose a risk as an opportunity might arise for subjects to discuss the format of the session and perhaps influence each other. However, during the session the researcher asked the subjects to give their own opinions on the quality of the subtitles and not to discuss their opinions with any other subjects they were acquainted with.

Now that we have dealt with the internal validity of the research study, we will move on to look at factors concerned with its external validity.

#### 3.3.4.2 External Validity

External validity refers to the extent to which the findings of a study can be generalised. Researchers (Frey et al. 1991 and Bailey 1994) identify three main factors that influence the degree of external validity: sampling, ecological validity and replication.

##### Sampling

This refers to the study of a small portion of the total population for the purpose of generalising about the entire population. It is closely related to subject selection discussed above. The two main distinctions within sampling are *random sampling* and *non-random sampling* (also known as probability sampling and non-probability sampling). Random sampling is designed to best guarantee a representative sample, reducing bias introduced by the researcher. However, it is not always possible to sample randomly from a population due, for example, to reasons of costs and difficulty finding complete population lists (Frey et al. ibid:134). Of the methods of sampling proposed by both authors (Frey et al. ibid:135, Bailey ibid:96), the two most relevant to this study are *purposive sampling* and *network sampling* (also called the *snowball technique*). Both of these methods of sampling are non-random. When purposive sampling is conducted the subjects are chosen non-randomly because they possess a particular characteristic pertinent to the study. In the current study the subjects were chosen on the basis of German being their mother tongue. Network sampling or the snowball technique is applied where subjects are asked to refer the researcher to other potential

subjects for the study. We implemented this technique when we contacted the potential subjects by email, thus allowing them to forward on this email or to ask another possible subject in person.

### Ecological Validity

This refers to the need to conduct research so that it actually reflects real-life situations, therefore increasing the possibility of generalising any research findings. Ecological validity is partly ensured through careful subject selection as already described.

A second important consideration is the experimental setting: Most of the time people watch a DVD on a television screen located in the sitting room of their own or someone else's home. Other people, however, prefer to watch a DVD on a laptop or mobile DVD player, for example when travelling. Armstrong et al. (2006c) asked the subjects in their study whether they had viewed a DVD on a laptop or a computer. All six subjects had watched a movie which had been downloaded from the Internet, on a PC or laptop computer. However, they did not express a strong interest in watching a bought or rented DVD on a PC or laptop. Based on these results we opted for the first option of viewing a DVD on a television instead of a laptop. Another point we need to raise about the "usual" setting of viewing a DVD is that it is probably more common for people to watch a DVD in pairs or in a group setting. However, in order to minimise the threat of intersubject bias, we conducted the evaluation sessions on an individual basis. If we had conducted the sessions with a few people simultaneously, this would also have complicated interviews and data collation. Likewise if we wanted to situate the evaluation sessions in a real-life setting, we would have to obtain permission from the subjects to conduct the sessions in their own living room. This, unfortunately, is not an option in this instance. Factors such as allowing ample time to travel to the various locations, costs associated with travel, scheduling sessions over the three-day and two-week periods, and accommodation arrangements[72] all have a negative impact on conducting these sessions outside an experimental situation.

Another factor that complicates the selection of an 'ideal' setting for the study relates to the gathering of data. As already indicated, in order to ensure that we gathered data on

[72] Students from outside of Dublin and who are studying at a university in Dublin usually live in shared accommodation on or off campus. Therefore conducting sessions in any of these locations leaves us exposed to some kind of unexpected disruption.

all clips, and based on experience from Armstrong et al. (2006c) we showed clips one by one, stopping to ask questions after each clip. This is an obvious compromise from an ecological validity point of view, as viewers normally watch a whole subtitled movie in one go, but one that was necessary for the reasons outlined above.

Frey et al. (ibid:137) point out that research can still be ecologically valid, even though it does not take place in real-life circumstances. Considering all of the factors associated with an ideal setting for this study, we created an experimental 'real-life' setting where the subjects could view the subtitles and at the same time avoid influencing factors such as those mentioned above. With this in mind, we implemented the following measures for the study to enhance the ecological validity of our research:

- The evaluation sessions took place in the Advanced Translation Research Lab (ATRL) in DCU, which is a room specially designed to replicate the average person's living room. The room has a wide-screen television and DVD player, together with comfortable seating and two lighting settings, since some viewers prefer the light settings to mimic those of a cinema auditorium while watching a DVD. The different light settings were shown to the subject before the session began
- The lab is located in a building the subjects are very familiar with
- The seating is such that the subject can position it freely in relation to the television screen as each subject has a different preference regarding the distance from the television
- There is a short introduction on the first clip with background music and noises to allow subjects to indicate if they want the sound level altered, before the first subtitle appears
- The mp3 recorder used to record the sessions is very compact and does not make any noise like a tape recorder might produce. This was to ensure that there was no additional noise in the room
- The lab is situated in a quiet area of the building that houses the School of Applied Language and Intercultural Studies, meaning there was no threat to any of the sessions from a sudden increase in noise levels associated with certain times of the day including lunch time and directly following the end of a lecture

### Replication

Tukey (1969 cited in Frey et al. 1991) argues that the results of a study must be repeated a number of times before the original study can be confirmed and extended. The results from one study simply cannot be relied on due to any number of the previously mentioned threats to internal and external validity. Replication means that a study must be repeated using the same procedures as the first study or by varying them in a systematic way. In this study we documented very clearly the procedures we followed, for example, how we obtained our texts, generated our subtitles using three different corpora, conducted the evaluation sessions, recruited our subjects and analysed our data to ensure the possibility of replicating the study and increasing the likelihood of external validity.

#### 3.3.4.3 Measurement Validity and Reliability

Measurement validity refers to the ability of a measurement technique to measure accurately what it is supposed to measure (Frey et al. ibid:119). For a measurement to be valid it must also be reliable. Measurement reliability requires a variable to be measured in a consistent and sound manner (ibid). The important difference between the two measurements is that validity is assessed at a conceptual level, while reliability is assessed at a numerical level. This study has drawn on MT evaluation literature to ensure we are measuring features of acceptability and intelligibility using established measurement techniques that have already been accepted as valid and are applicable to this research area.

Measurement reliability indicates the amount of error associated with a measurement. A measurement always contains a *true score component* and an *error score component*. The error score component can be divided into *random error* (when subjects make mistakes) and *measurement error* (when researchers make mistakes). The reliability of a measurement technique is certainly increased if the measurement error is reduced. Measurement error is unavoidable, but there are particular measures that can be implemented to reduce its extent. Frey et al. (ibid:120) propose pilot testing, questionnaires, interviews and observations as methods for reducing occurrences of errors. Regardless of the method used to increase the validity and reliability of the research, Frey et al. point out that the particular method used must fit the concept being measured (ibid:124). They also make the point that measurement validity is commonly

increased with the use of qualitative measures and in contrast, measurement reliability is commonly increased with the use of quantitative measures. This idea justifies our approach of using both types of measures.

## 3.4 Practical Research Design

After discussing the theoretical elements of the study, we now need to implement these in a practical research design. The research design includes general topics of corpus compilation, subject selection, text selection, interview questionnaire design, and details on how each phase of the methodological framework was implemented to gather data.

### 3.4.1 Corpus compilation

Given the findings from the pilot study (Armstrong et al. 2006c) we used only segment-aligned English and German subtitles to train the EBMT system. We noted earlier (see section 1.5.6) that due to time constraints Armstrong et al. (ibid) were unable to run quality checks on the subtitles used for the training corpora. However, for the current study we conducted evaluation sessions of the commercial DVD subtitles used in the training corpora. Based on judgements of three native German speakers, the subtitles taken from the commercial DVDs were deemed acceptable (see Appendix A for results). As we have already mentioned in section 3.1.1, for the pilot study we had used broad definitions of 'heterogeneous' and 'homogeneous' data. In the current study we wanted to pursue the notion of homogeneous data further to investigate the degree of homogeneity necessary to generate the most acceptable set of DVD subtitles. As a result we trained the EBMT system on three different corpora that become progressively bigger and less genre-specific and which are outlined below.

#### Three Corpora for the Current Study

The first corpus is a bilingually-aligned corpus of subtitles taken from the first four *Harry Potter* movies in the series that is available to view on DVD.[73] These include *Harry Potter and the Philosopher's Stone* (2001), *Harry Potter and the Chamber of Secrets* (2002), *Harry Potter and the Prisoner of Azkaban* (2004), and *Harry Potter and the Goblet of Fire* (2005). We call this the Harry Potter Corpus (HPC) or **Corpus A** and it belongs to the fantasy genre. Corpus A contains approximately 7,000 subtitles.

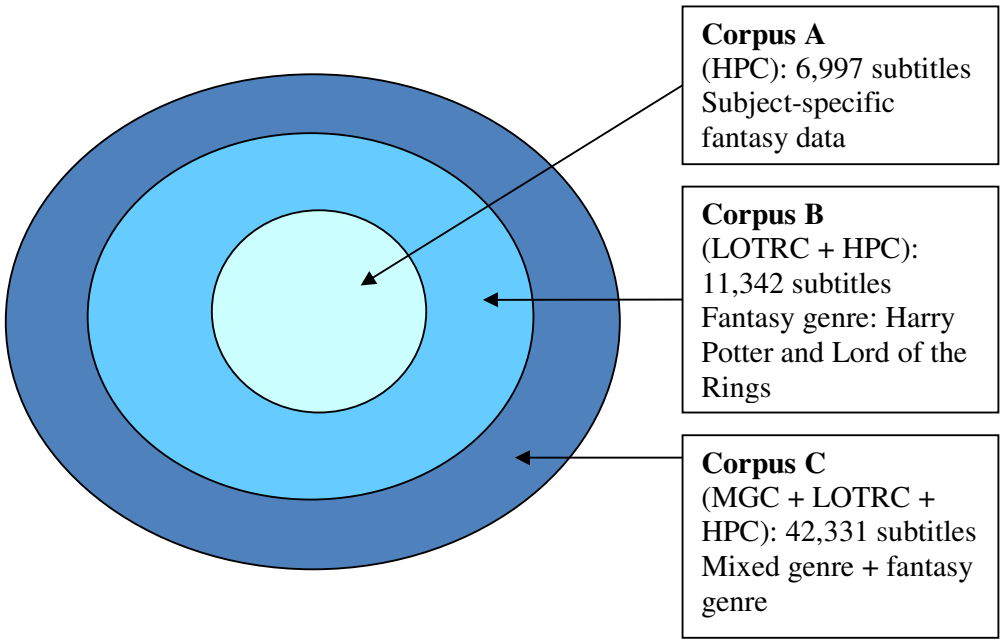
[73] These were the only Harry Potter DVDs available at the time of compiling the corpora (January 2007).

The second corpus, **Corpus B**, is a combination of two corpora: a bilingually-aligned corpus of subtitles taken from the *Lord of the Rings* trilogy, called the Lord of the Rings Corpus (LOTRC) and Corpus A. Like Corpus A, Corpus B belongs to the fantasy genre. However, the genre has been expanded slightly to include two different kinds of fantasy movies. Corpus B contains approximately 11,400 subtitles.

The third corpus, **Corpus C**, is also a combination of two corpora: a bilingually-aligned corpus of subtitles taken from twenty-five movies on DVD from various genres ranging from action/adventure to romance and period dramas, called the Mixed Genre Corpus (MGC) and Corpus B. Corpus C contains approximately 42,300 subtitles.

Creating three separate corpora allowed us to build on work already conducted within the MovRat project (cf. Armstrong et al. 2006c) relating to factors which might have an influence on the acceptability of EBMT-generated subtitles, notably genre and size of the corpora. In the current study we thus distinguished between subtitles which are very subject specific and from the fantasy movie genre (HPC), those which were slightly less subject specific, but still remained within the same movie genre (LOTRC), and those which were not considered subject specific, came from an array of movie genres, and which contained what might be considered 'general, everyday' language (MGC). Figure 3.3 overleaf is a visual representation of the three corpora used in the current study.

**Figure 3.3:** Visual representation of the three corpora used in the current study




Although our corpora are relatively small by contemporary standards, we have been encouraged by the work of Estrella et al. (2007), which highlights the fact that often the need for 'large' amounts of texts for testing and training is overestimated and reliable evaluation results, both human and automatic can be obtained by using fewer documents than expected, thus reducing evaluation costs (time and effort) (ibid:167). Our work uses small test sets compared to the usual test sets used in MT evaluation campaigns. For example, CESTA[74] used 790 segments for their test set, and groups participating in the IWSLT-08[75] shared task used on average 1,000 sentences in their test set. We use 178 subtitles (segments) in our test set overall (which is broken down into six movie clips of between 23 to 36 subtitles per clip). The use of smaller test sets is understandable given that we use qualitative and human evaluation techniques in this study, and not solely automatic metrics to evaluate the intelligibility and acceptability of the MT output.

[74] <http://www.technolangue.net/IMG/pdf/Rapport_final_CESTA_v1.04.pdf> [Accessed 10 March 2009].
[75] <http://www.slc.atr.jp/IWSLT2008/archives/2008/10/references.html> [Accessed 10 March 2009].

The process of compiling the three corpora followed the same steps as reported in Armstrong et al. (2006c). In the current study we were more specific regarding the films we chose for the corpus, focusing on the fantasy genre and movies from sequels. There are two differences between the corpus compilation stage reported on in Armstrong et al. (ibid) and the current study. The first was that for this study the researcher had to compile the three corpora without any help from the other researchers, but could build on work already completed during the pilot study. This involved extracting subtitles and running Perl-encoded scripts on the corpora to clean them up. Fortunately we were able to hire three students to manually check the alignments, thus speeding up the process and this allowed the researcher to work on other aspects of the current study.[76] Overall the corpus compilation stage could still be considered time consuming, as previous efforts of aligning segments using Perl scripts proved to be consistently erroneous. We therefore preferred to align the subtitle segments manually. We were confident that we had correctly aligned the subtitles at segment level, and we gained a good knowledge of the content of the corpus. These two points were beneficial for the stages that followed. The second difference concerned the quality of the German human-translated subtitles that were included in the training corpus. We previously mentioned for the pilot study that we attempted to assure the quality of the subtitles by only taking subtitles from major motion pictures which tend to have high-quality human subtitles. The issue of quality was not pursued further during the pilot study due to strict time constraints. However, for the current study we wanted to revisit this point and we conducted a small-scale study with German native speakers, which we mentioned earlier.

#### 3.4.1.1 Ripping DVDs for Research

In the previous sections on corpus compilation we described how we extracted the subtitles from the DVDs. The process of extracting videos or music and the associated subtitles from a DVD, CD, VHS tape or vinyl record to a computer hard drive is known as 'ripping' (Huang 2007:131). In the following sections we address the legal issues surrounding this practice, given that it was central to the present study.

Recent developments in digital technology and especially in copyright laws both in the USA and the EU have given rise to technical and legal difficulties in using DVDs for

[76] It should be noted here that even though the alignments were manually checked by three student translators, who had an excellent knowledge of English and German, the researcher always carried out a final check of the aligned segments.

pedagogical purposes. Huang (ibid:129) maintains that nowadays, under the 2001 EU Copyright Directive (EUCD),[77] teachers who rip a DVD for teaching purposes could face a jail sentence for breach of copyright. Issues regarding the copyright of subtitles are equally important to researchers. The two issues we need to highlight are availing of subtitles from commercial DVDs and putting new subtitles on a selection of movie clips taken from commercial DVDs to show to a selected audience. In most common law countries (including Ireland) we have a doctrine of limitations and exceptions to copyright known as 'fair dealing'. This doctrine is a list of categories that could be considered possible defences against an action for violating any rights of copyright. The doctrine of fair dealing is similar to the doctrine of fair use implemented in the US; however, fair dealing is not as flexible as the corresponding term used in the States and it cannot be applied to any act that does not fall within one of the stated categories. In Ireland Chapter 6 of the Copyright and Related Rights Act, 2000, lists the acts permitted in relation to works protected by copyright. In particular, act 50 deals with fair dealing in relation to research or private study. Parts (1) and (4) of this act are most relevant to our study (Copyright and Related Rights Act 2000):

> (1) Fair dealing with a literary, dramatic, musical or artistic work, sound recording, film, broadcast, cable programme, or non-electronic original database, for the purposes of research or private study, shall not infringe any copyright in the work.
>
> (4) In this Part, "fair dealing" means the making use of a literary, dramatic, musical or artistic work, film, sound recording, broadcast, cable programme, non-electronic original database or typographical arrangement of a published edition which has already been lawfully made available to the public, for a purpose and to an extent which will not unreasonably prejudice the interests of the owner of the copyright.

These two parts of the act deal with our research and clearly outline the allowed use of a movie and its related characteristics, including subtitles. It is not clear from the act whether it differentiates between commercial and non-commercial research, as commercial research could be considered an infringement of copyright work. The current study is of a non-commercial nature dealing with MT evaluation methods.

[77] EU and USA copyright laws are similar in many ways so we will just mention the EU law in this instance.

### 3.4.2 Subject Selection

We have previously outlined the criteria we considered when recruiting subjects (3.3.4). A similar study to ours, reported in Volk (2008) and Volk & Harder (2007), evaluates MT output using only automatic metrics to calculate 'quality' scores and does not gather end-user subjective opinions.[78] Therefore their studies are not subjected to the threat of subject characteristics and skills influencing the scores. Popowich et al. (2000) is the only study prior to the work of Armstrong et al. (2006c) that used a human evaluator to evaluate the quality of automatically-translated subtitles. In this instance one translator was used to evaluate the quality of the MT output. To our knowledge, the work of Armstrong et al. (ibid) and the current study are the first to use end-users to evaluate automatically-generated subtitles, meaning there is no literature on the type of subjects who would minimise any bias on the results or indeed on the ideal number of subjects for each kind of session. We previously outlined the kinds of evaluation subjects were recruited for during the pilot study and how the subjects were recruited. The following paragraphs detail how the subjects were chosen for the prospective and retrospective phases of the current study.

#### Prospective phase

The prospective phase incorporated one evaluation phase:

- Human evaluation of the reusability of TL subtitles

For this phase we recruited three members of staff in the School of Applied Language and Intercultural Studies at Dublin City University. All three were German native speakers and were fluent in English. They lecture or have lectured in German and have similar educational backgrounds. There were one male and two females in the group. None of the subjects in this group had any previous knowledge of the research being conducted in this study. The reasons for recruiting members of staff included their language and linguistic expertise and availability. The sessions conducted with these three subjects relied heavily on being able to give subjective opinions on linguistic-based tasks, requiring subjects to make judgements on the reusability of language segments. The number of subjects participating in these sessions was still in line with research mentioned in section 1.5.6.

[78] In this context they did not gather opinions from viewers of the subtitles, as is the case for the current study. However, they did gather opinions from professional subtitlers, who were the 'end-users' of the system outlined in Volk & Harder (2007) and Volk (2008).

### Retrospective phase

Subject selection criteria have already been outlined in section 3.3.4. In this section we focus on the number of subjects appropriate to the end-user evaluation, and on actual subject selection. From the previous discussion on this point, Arnold et al. (1994) note that bigger groups make the results more reliable, but how big is big enough in this context? Chiaro (2006) says we must not forget that a survey is not a census; "in other words, a survey does not set out to assess the *whole population* but just a representative sample of the population." She points out that students involved in exploratory field work have strict budget constraints and tight deadlines, two factors which make it unrealistic for them to work with very large sample sizes. She outlines that "for a tentative investigation, according to mathematicians, 30 seems to be a magic number". This small, manageable group still allows for significant results to be drawn if the correct statistical test is used, specific to small groups. Chiaro (ibid) notes that the most important thing to remember when using a sample in and around the 30 mark is that:

> [As] long as we are aware of the fact that our study is a small one and that it is *exploratory* in nature and not the final word on the matter, such hypothetical studies would be perfectly respectable examples of quantitative research.

With these comments in mind, we aimed for at least thirty in our group. We worked on the basis that the more subjects we could recruit over a two-week period, the more data we could generate and therefore the more reliable statistical tests we could conduct on the data. We scheduled the evaluation sessions for the beginning of June, following the second semester exams. We advertised the evaluation sessions by email (see copy included in Appendix B), contacting only students from the School of Applied Language and Intercultural studies, to facilitate the 'snowball technique' (see section 3.3.4.2). Unfortunately we had not anticipated that so many of the German-speaking students would have left Ireland so soon after the exams, and therefore we received a very low response from the students. For the June sessions we recruited twelve subjects, all of whom were postgraduate students, either pursuing a taught postgraduate course or a PhD by research. The sessions in June 2007 were conducted over a three-day period. We knew that we could not increase the subject number from twelve during the summer months and instead opted to wait until the beginning of the next academic year (September 2007) to begin our recruitment phase once again. This time we conducted

the evaluation sessions at the beginning of the academic year, and advertised the evaluation sessions campus-wide. Class lists for each module being taught at DCU are available online and if a class consists of students participating in a study-abroad programme, this is marked by an X following the module code.

When recruiting for the second set of sessions we were able to target groups from abroad, and therefore more likely to reach students from German-speaking countries. Like with the first set of sessions, we advertised this set by email, and offered modest remuneration to students for taking part on the basis that such material incentives would increase participation (cf. Göritz 2006).

We thereby recruited a further thirty-two subjects for the end-user evaluation sessions (November 2007). That meant, in total, we had forty-four participants for the retrospective evaluation sessions. We will note here one point regarding the validity of conducting two separate evaluation sessions. Since we conducted both evaluation sessions in the same manner and used exactly the same material on both occasions we knew that the sessions were not compromised in any way and that the data from the first session would not be contaminated or redundant for use with the data collated during the second set of sessions.

### 3.4.3 Text Selection

Decisions regarding test data subtitles for the end-user evaluation sessions were based on experience outlined in Armstrong et al. (2006c) and the work conducted during the prospective phase. Armstrong et al. (ibid) showed that the subtitled *Harry Potter* clips were more positively received than the subtitled clips from *The Bourne Identity*. Keeping in line with our previous comments on this matter (cf. ibid:178), this could be due to the fast-moving pace of an action movie such as *The Bourne Identity*, which includes many shot changes in contrast to the slower, calmer pace of the fantasy movie *Harry Potter*, which is aimed at a younger audience. Based on these findings we opted to subtitle only movie clips from a *Harry Potter* movie. This allowed us to narrow our study to subtitles from a particular genre, using corpora that contained different levels of SL repetitions, different number of subtitles and varying levels of homogeneity, and to test the relationship between these variables and the intelligibility and acceptability of machine-generated subtitles.

The next step was to choose the exact clips we would re-subtitle from a *Harry Potter* movie, as it is not feasible in this study to re-subtitle and ask subjects to view the entire movie (see section 3.3.4). The choice of specific subtitles began during the corpus-analysis stage (prospective phase). For no reason other than it is the first film in the sequel of *Harry Potter* movies, we chose clips from *Harry Potter and the Philosopher's Stone.* The specific clips were chosen by locating clusters of SL repetitions and from these we chose the 'best' ten clips, in other words the clips that showed the highest number of SL repetitions. The methods used to locate the SL repetitions are outlined in the discussion of the prospective phase (see section 4.1).

As already indicated three German native-speaking lecturers then evaluated these ten clips with a view to ascertaining how reusable the EBMT-generated TL subtitles they contained were. This was also done during the prospective phase. Based on the data collected from the subjects during this phase, the six clips with the potentially 'most reusable' TL segments were used as the movie clips for the retrospective end-user evaluation sessions.

A final word should be said here about why we chose to re-subtitle clips from a movie that already contained human-translated subtitles on the DVD. In order to calculate automatic metric scores of the EBMT output, a reference translation is required against which MT output can be judged. This means that if we want to calculate a score for each of the subtitle translations we produced using the EBMT system then we have to use a source text or subtitle that already has an existing target text or subtitle translation. Therefore we re-subtitled clips from the first *Harry Potter* movie, instead of perhaps subtitling bonus material from the movie, which is supplied only in English.[79]

### 3.4.4 Interview Questionnaire Design

For this study we used an interview questionnaire (Oppenheim 1992) to gather opinions from subjects. Literature on qualitative interviews in the traditional sense advises that interviewers should take an interview guide, outline or schedule with them when conducting an interview. Rapley (2004:17, author's italics) points out that:

[79] The idea, however, of subtitling such bonus material is one of the main areas of focus of the automation of subtitles and is certainly an area for future research. In many cases the bonus material is only supplied in the language of the original soundtrack, and it is not usually included in a dialogue list.

> The actual content of the list of questions is *initially* generated in negotiation with the relevant academic and non-academic literature, alongside your thoughts and hunches about what areas *might* be important to cover in the interview.

This point also applies to the questions administered during our interview. However, in contrast to the traditional interview, the questions in this instance are more structured and in some ways 'fixed' within the scope of the questionnaire which contrasts with the usual interview guide where the questions can change depending on the interview. Additional comments are nonetheless very welcome from subjects and these are analysed under subsequent headings from the more fixed sections.

Armstrong et al. (2006c) detailed how they tested a range of questions relating to attitudes to subtitles and translation technology as well as opinions on EBMT-generated output, without asking a specific set of questions relating to any particular movie clips. We adopted the same format for the current study, for example, beginning with general background information and then focusing on particular areas of interest. However, we narrowed the focus of the questions in order to elicit subjective opinions on the intelligibility and acceptability of EBMT-generated subtitles offered on specific movie clips, and then combined this information for analysis purposes to gain a more holistic insight into subjects' opinions.

In general a questionnaire can begin with either factual questions or attitude questions, depending on the type of questionnaire, but whatever the researcher decides it is important to "avoid putting ideas into the respondents' mind or to suggest that they should have attitudes when they have none" (Oppenheim ibid:112). This point is important if we need spontaneous response on the same points later on. The sequence of the questions must also be attractive and interesting to the subject in order to gather the most reliable information. Intimidating the subjects early on or researchers making their own attitudes very obvious will result in a bad rapport between the researcher and the subject (Ackroyd & Hughes 1992, Rapley 2004, Weiss 1994).

A common approach to question sequence is the *funnel approach*, followed by various *filter* questions (Oppenheim ibid:110). The approach refers to a sequence that begins with a very broad question and then gradually narrows down the scope of the questions to eventually become a very specific question. It is also possible to begin with a filter

question, used by researchers to exclude a subject from a particular question sequence if the sequence is irrelevant to that subject.

### Questionnaire: Prospective Phase

Within the prospective phase we used a type of self-administered cum interview questionnaire to collect opinions on the reusability of TL subtitles in different contexts. This involved two sets of booklets: Clips and Options.

### Clips Booklets

Describing the questionnaire administered to the three subjects in this phase as a type of self-administered cum interview questionnaire means that the subjects wrote in answers to questions in an individual booklet provided by the researcher, while the researcher also filled out an identical booklet. This allowed the researcher to take notes on a particular subtitle or observation subjects might have made. For this session each subject received two sets of booklets. The first set of ten booklets (Clips Booklets 1-10) contained:

- The context in which the clip is set
- The original English subtitle
- The speaker of the subtitle
- And - in cases where there were repeated SL subtitles - three translations of the original subtitle, each chosen by the EBMT system depending on which of the three training corpora was used (see sample copy included in Appendix C)

### Options Booklets

The second set of ten booklets (Options Booklets 1-10) contained:

- Alternative translations for the repeated SL segments, where these alternatives, although extracted from the corpora, had not been chosen by the EBMT system (see sample copy included in Appendix D)

In the Clips Booklets, subjects were asked to indicate whether or not the subtitle chosen by the EBMT system was deemed acceptable or not (yes, no or don't know) in the particular context outlined. If the subtitle was not repeated (internally or externally) within the corpora, no EBMT subtitle was provided, as we wanted to focus on translations of repetitions in this instance. There were 61 sets of subtitles presented,

with 3 subtitles per set. In the Options Booklets subjects were asked to indicate whether or not the alternative translation(s) was (were) deemed acceptable in the particular context. There were 40 sets of alternative translations presented, with anywhere between 2 and 12 subtitles per set.

For both sets of booklets there were three options in response to whether a segment is acceptable: 'yes', 'no' and 'don't know'. This goes against a warning given by De Vaus (2002:106) claiming that "the danger with using 'don't know' and 'no opinion' alternatives is that some respondents select them out of *laziness*". However, as we implement an interview questionnaire rather than simply a self-administered, group-administered or mail questionnaire, the researcher can ensure that there is a valid reason for the subject to choose the 'don't know' option. During the sessions there was a lot of dialogue between the subject and the researcher, allowing the researcher to ensure questions were answered fully. In addition both sets of booklets allowed subjects to elaborate on answers if they chose to do so.

### Questionnaire: Retrospective Phase

During the retrospective end-user evaluation sessions we used an interview questionnaire to gather human judgements on the intelligibility and acceptability of machine-generated subtitles (see copy included in Appendix E). The design of the questionnaire builds on the work of Armstrong et al. (2006c). The questionnaire began with three background sections: the first elicited information on the subjects, including reference to educational background and university standing (3 questions); the second elicited information relating to *Harry Potter* (7 – 14 questions, depending on the answer to the filter questions); and the third elicited information on subjects' watching of subtitled movies on DVD (10 questions). These were followed by a set of nineteen questions (a mix of open and closed) which were asked after each of the six movie clips. Four of the twelve closed questions in the background sections are followed by an open question allowing the subject to elaborate on a particular answer. For each clip we included a filter question to establish if subjects understood the movie clip, followed by two internal checks, to make sure subjects did not simply feel obliged to say they had understood the clip. Other questions ask if they fully understood the clip using only the subtitles or a combination of subtitles and/or image and/or soundtrack. We aimed to establish how helpful the subtitles were in the understanding process (Example 3.4).

**Example 3.4:** Questions used to establish if the subjects comprehended the subtitles

1 After watching this clip did you understand what was happening in the clip?

a) Yes

b) No

Comments:

**Internal checks**

2 When Melvin picks up the dog to talk to him, why is the woman looking over at them impressed?

3 What are the two women discussing, who are sitting on the couch?

Armstrong et al.'s (2006c) research tested some questions during the face-to-face and online evaluation sessions which are also used in this study either in the same format or which have been developed further. New questions introduced into the questionnaire relating to the four quality characteristics were informed by FEMTI and literature on recipient evaluations (Trujillo 1999, Bowker & Ehgoetz 2007). We detail the questions specific to each of the quality characteristics in Tables 3.1 – 3.4 below:

**Table 3.1:** Questions used to elicit responses on the comprehensibility of subtitles

| **Comprehensibility** |
|---|
| After watching the clip, did you understand what was happening? |
| Two questions asked as internal checks specifically about the characters in the clip |
| On a scale of 1-6 (6 being very comprehensible, 1 being incomprehensible) where would you locate the subtitles for this clip? |
| What did you use to understand the clip (Soundtrack, Image, Subtitles)? |

**Table 3.2:** Questions used to elicit responses on the readability of subtitles

| **Readability** |
|---|
| Was the speed of the subtitles suitable? |
| Did you notice any subtitles which seemed out of context? |

**Table 3.3:** Questions used to elicit responses on the style of subtitles

| **Style** |
|---|
| On a scale of 1-6 (6 being appropriate style, 1 being inappropriate style) where would you locate the subtitles for this clip? |
| Did anything in the subtitles during this clip particularly bother you? (+Examples) |
| Did anything in the subtitles during this clip particularly amuse you? (+Examples) |

**Table 3.4:** Questions used to elicit responses on the well-formedness of subtitles

| **Well-formedness** |
|---|
| Did you notice errors in the subtitles (what kind)? |
| On a scale of 1-6 (6 being not annoying at all, 1 being very annoying), where would you put the errors? |
| Are the subtitles acceptable for viewers who would not understand the soundtrack? |
| Are there any particularly well-translated subtitles? |

Some of the questions outlined above could produce related answers, for example a question indicating if a subject noticed any errors (well-formedness) and a question about whether subjects noticed anything about the subtitles that bothered them (style). The inclusion of somewhat related questions in different categories allowed us to check for intra-subject agreement (qualitatively) and it adds to the reliability and validity of the responses. Within the context of an interview it is normal for subjects to repeat points they consider very important or to repeat a point in one category that may also fit into a different category. This also means that subjects might have additional comments

they did not think of the first time they were asked a question about the subtitles, but recalled some information triggered by a different question.

The questionnaire finished with an overall section comprising four questions given in Table 3.5 below. One asked if subjects noticed any repeated subtitles; one asked if they considered the subtitles on the clips with the known soundtrack more acceptable than the subtitles on the clips with the unknown soundtrack, and two asked subjects to comment on their satisfaction (rating scale and open question). The question on repeated subtitles allowed us to investigate whether the detection/perception of repeated subtitles has a negative effect on viewers' satisfaction, and the question regarding the soundtrack language allowed us to investigate whether a known versus an unknown soundtrack language influenced judgements on the four quality characteristics.

**Table 3.5:** Questions used to elicit overall satisfaction with the subtitles

| |
|---|
| Did you notice any repeated subtitles throughout the clips? |
| Are the subtitles more acceptable on clips with a Dutch language soundtrack or with an English language soundtrack? |
| On a scale of 1-6 (6 being very satisfied, 1 being very dissatisfied), where would you rate the subtitles overall? |
| Do you have any overall comments regarding your satisfaction/dissatisfaction with the subtitles or the use of EBMT to generate them? |

These four questions within the *overall* category were included to ask the subjects to reflect holistically on the subtitles they just viewed. As we discussed earlier one area we want to investigate is whether technology, such as MT, would be of benefit when subtitling movies within a series, such as *Harry Potter* for example. Practitioners in the subtitling industry (Languages and the Media Conference 2006: panel discussion) have raised the point that if technology is used, the subtitles will seem very repetitive like technical manual translations.

We also introduce an additional question in this category. During the evaluation sessions subjects were asked one question after each of the three clips with the unknown language soundtrack (Dutch):

- Would you use these subtitles on a DVD if you did not understand the soundtrack language?

This individual question is considered in the *overall* category as it is not specific to any of the four quality characteristics, but rather to the overall acceptability of the subtitles when the viewer does not understand the soundtrack. It aims to investigate the popular claim that if a viewer has knowledge of the soundtrack language, they will be more critical in their evaluation of the machine-generated subtitles (cf. Armstrong et al. 2006c).

In the preceding sections we discussed the theoretical and practical elements of the research design. We now move on to outline the practical implementation of the elements previously discussed. As already indicated, there are two phases implemented in this study (cf. Figure 3.1):

- Prospective Phase
    - Corpus analysis
    - Human evaluation of the reusability of TL subtitles
- Retrospective Phase
    - End-user evaluations

### 3.4.5 Prospective Phase

#### Corpus analysis

Using SDL Trados' Analyse tool and Microsoft Word, the first step in corpus analysis was the counting and then identification of SL repetitions in the corpus and the corresponding TL translations. These data were then used to investigate the potential reusability of given TL segments in different contexts, as simply calculating the number of SL repetitions does not give us enough information regarding the usefulness of the TL segments in an EBMT environment. During the corpus-analysis stage we gathered statistics on the three corpora, which can be used at a later stage to examine whether a relationship exists between corpus size, number of SL repetitions and homogeneity of the corpus, and the intelligibility and acceptability of the subtitles from the point of view of the end-users.

### Human evaluation of the reusability of TL subtitles

This stage has two aims: firstly, to analyse the quantitative data generated in the previous section from a qualitative viewpoint and to identify the corpus that is considered to contain the highest number of reusable TL segments in different contexts; secondly, using the information about reusable TL segments we can choose the six movie clips[80] out of the ten presented in this section to use for the retrospective phase of the study. Building on work by Flanagan & Kenny (2007), we describe the procedures implemented to evaluate the reusability of subtitle translations in different contexts and to select the most suitable movie clips for our evaluation sessions.

As previously outlined in section 3.4.2 three subjects participated in this prospective evaluation stage and they were presented with two sets of subtitle booklets (10 Clips Booklets and 10 Options Booklets), instead of being shown subtitled movie clips on a television screen. The subtitles provided in these booklets were from the ten *Harry Potter and the Philosopher's Stone* movie clips described earlier. We have previously outlined the contents of the booklets. The ten Clips Booklets were used to gather subjective opinions of three subjects and we use these results to ascertain which movie clips contained the highest number of acceptable EBMT translated subtitles in a given context. These data were one of the criteria used to select the six clips for the retrospective phase. The second criterion used to rank the clips was the number of different EBMT-generated subtitles each clip contained.[81] The ten Options Booklets were used to present the subjects with alternative translations of the subtitles in the Clips Booklets. These alternative translations were present in the corpora but were not selected by the EBMT system.[82] This allowed us to investigate our quantitative data

[80] We used only six of the ten clips for the retrospective phase, as otherwise the evaluation sessions would have lasted too long, possibly threatening the validity of the study.

[81] If we did not specify different EBMT-generated subtitles (subtitle types), and counted every subtitle (subtitle tokens), this would skew our results. An example of this is clip 2 that contains 6 EBMT-generated repeated subtitles (tokens), however, 4 of these subtitles are exactly the same, and therefore we say it contains only 3 repeated subtitles (types) (see Table 4.2).

[82] The number of times a source segment and the same target segment translation is present in a corpus has an impact on the likelihood of the target segment being chosen as the most suitable translation. This is because the number of times a source and target segment occurs in the corpus has an impact on the word alignment. GIZA++ is used to train the word alignment, with GIZA++ being based on the co-occurrence and frequency of the source word and target word. Therefore, the number of times a source and target segment is repeated will impact on the alignment result and alignment probability, and consequently the resulting segment translation (Jinhua Du: personal communication).

further and to make some claims regarding the reusability of repetitions identified in each corpus.

The sessions were structured in the following way. Firstly, subjects were asked to read a pre-viewing briefing and to sign the briefing to give their informed consent (see copy included in Appendix F). By doing so they also agreed to the session being recorded on cassette tape, to capture any additional comments they may have made during the session.[83] Next, the participants were asked to look at the set of Clips Booklets and beginning with Clip 1, to read the context given for the clip. They then read through each English language subtitle in order, and referred to the EBMT-chosen translated subtitle if available, noting whether or not the translated subtitle was acceptable in the given context (ticking yes, no or don't know box). After evaluating each EBMT-chosen subtitle subjects referred to the Options Booklet to see if there were any additional translations offered in the corpora. In some cases there were no alternative translations offered. In the cases where there were other options to choose from, the subjects were once again asked to indicate whether or not they thought the subtitle was acceptable in the context stated (again, yes, no and don't know options). The session continued in this manner until all ten Clips and Options Booklets were completed. Each session described here last approximately one hour. The results from the prospective phase are presented in Chapter 4.

### 3.4.6 Retrospective Phase

The aim of the retrospective phase is to measure the intelligibility and acceptability of EBMT-generated subtitles from the point of view of the end-users, and this is done by conducting individual evaluation sessions with the end-users.

This phase consisted of forty-four individual evaluation sessions, which took place in the ATRL lab. Each session lasted between one hour and one hour fifteen minutes, depending on the length of time the subject spent discussing the various topics or if they wanted to add additional comments at the end of the session, and only the researcher and the subject were present during the session. The session began with the researcher reading a pre-viewing briefing, outlining the role of the subject and the format of the

[83] Unfortunately during these sessions there was no mp3 recorder available, and therefore we had to record the sessions on a cassette tape.

session (see copy included in Appendix G). The subject signed the briefing to confirm he/she agreed with the details of the session and understood how the data generated were going to be used. They also gave their full consent for the data to be used in this study. The session was recorded on an mp3 recorder to capture any additional information the researcher did not record in writing. The researcher began each session by asking questions from the interview questionnaire, starting with the background sections as outlined in section 3.4.4. The subject then viewed the first clip. After viewing the clip the subject answered questions specific to the clip. The same procedure was repeated until all six clips had been viewed. All answers were recorded by the researcher on the questionnaire and on the mp3 recorder. Four of the clips lasted two minutes and two of the clips lasted four minutes. The results of this phase are presented in Chapter 5.

## 3.5 Concluding Remarks

This chapter outlined the methodology used in this study. The chapter began by outlining the research questions and the research design. It then described in detail the EBMT system used to generate the subtitles for this study. The chapter then moved on to look at the theoretical aspects of the research design on the one hand, and the practical aspects on the other. Theoretical concerns include variable relationships, operationalisation, units of analysis, and internal and external validity of the study. The practical issues relevant to this study included the compilation of the corpora, subject and text selection and the design of our data collection method, namely the interview questionnaire. The final section illustrated how we implemented the methodological framework. In addition, this chapter highlighted how each stage of the methodology added to the current literature, given that corpus-analysis research in the area of MT is not usual practice and apart from the work of Armstrong et al. (2006c), automatically-generated subtitles have not been evaluated by end-users in any of the relevant published literature. We now move on to Chapter 4, which presents and analyses the data generated during the prospective phase (corpus analysis and human evaluation of the reusability of TL subtitles), followed by Chapter 5, which presents the data generated during the retrospective phase (end-user evaluation sessions).

# - Chapter 4
# - Prospective Phase: Results and Analysis

# 4 Prospective Phase: Results and Analysis

This chapter is divided into two sections: the first (4.1) presents the results from the corpus-analysis stage, and the second (4.2) presents the results from the human evaluation of the reusability of target language subtitles. We then discuss the relationship between the results from both stages.

## 4.1 Corpus Analysis

The following sections describe the practical implementation of the corpus-based study and once again build on work reported in Flanagan & Kenny (2007). Firstly in order to profile the training corpora, we identify SL repetitions. Therefore we started the corpus-analysis stage by carrying out some simple repetition and match analysis using SDL Trados Translator's Workbench 'Analyse' function with the three basic corpora, HPC, LOTRC and MGC. Because we subtitled movie clips from the first *Harry Potter* film, we wanted to get a rough idea of (a) how many repeated segments exist within the entire HPC (**internal repetitions**), (b) the extent to which segments in HPC recur in exactly the same form in the LOTRC on the one hand, and the MGC on the other (**external repetitions**), and (c) how many segments are repeated both within the HPC and externally to the HPC (**both internal and external repetitions**). In Table 4.1 below an internal repetition refers to a repetition that occurs when the segment is repeated only in the current corpus. An external repetition refers to a repetition that occurs when a segment in the current corpus is repeated only in one or more other corpora, but not in the current corpus. To find out (a) we simply analyse the HPC against an empty Translation Memory (TM); to find out (b) we analyse the HPC first against a TM containing all the source and target segments from the LOTRC, and then against a TM containing all the source and target segments from the MGC; and to find out (c) we add (a) and (b) results together. By comparing the HPC against already seeded TMs, we get a score for the number of **exact matches** with that TM. An exact match is when the segment in the text being analysed is a 100% match with a segment contained in the TM. For our purposes 'external repetition' and '100% match' or 'exact match' are co-terminous.

**Table 4.1:** Repetition rates and 100% matches between HPC and three different TMs

| Corpus | Translation Memory | Type of match | Repetitions/100% Match | % of HPC |
|---|---|---|---|---|
| HPC | Empty | Internal | 1181 (Repetitions) | 17 % |
| HPC | LOTRC | External | 598 (100% match) | 13 % |
| HPC | MGC | External | 1150 (100% match) | 16 % |
| HPC | Empty + LOTRC + MGC | Internal + External | 2929 (repetitions + 100% matches | 42 % |

Table 4.1 shows the HPC has 1181 internal repetitions, 598 100% matches with the LOTRC, 1150 100% matches with the MGC and 2929 internal and external repetitions. This is an indication that there are SL segments repeated within the HPC and also SL segments in both the LOTRC and MGC which are the same as segments in the HPC, and therefore could potentially provide good translations for the corresponding HPC segments.

After looking at general statistics for the different corpora, we wanted to choose ten movie clips from the first Harry Potter film, *Harry Potter and the Philosopher's Stone* (2001) which would be used in the next stage (human evaluation of the reusability of TL subtitles). Given the fact that we wanted to focus on repetitions and the reusability of their translations, it was decided to locate clips which showed high levels of repetition. This way there was a larger number of translation examples to show the subjects during the prospective phase, and more data to work with when trying to establish reusability of the subtitles in different contexts. The approach for selecting clips with the highest number of repetitions may seem somewhat biased, but in order to cast most light on the research questions posed in this study, it was important to use clips that exhibited high numbers of repetitions.

Even though technology such as SDL Trados Translator's Workbench can provide quantitative data on the contents of a corpus very quickly, it is necessary to manually go through the data, in order to find out exactly where the repetitions occur (relative to the clips), and we did this using a colour-coding scheme (see Figure 4.1 below). Repetitions

that occurred only within the first *Harry Potter* movie, *Harry Potter and the Philosopher's Stone* (2001), (internal repetitions) were marked in yellow; repetitions that occurred external to the first movie were marked in red (either in any of the other three *Harry Potter* movies and/or LOTRC and/or MGC); segments that were repeated both internally to the first *Harry Potter* movie as well as externally in the rest of the HPC, and/or the LOTRC and/or the MGC were marked in green.

We used Microsoft Word as our editing environment. This allowed us to group together all repeated segments within a corpus (using Word's Sort function), and thus identify exactly which segments accounted for the repetitions counted by the SDL Trados's Analyse Tool. Microsoft Word also gave us a convenient way of colour-coding segments. Once the coding was done, if we selected a clip which had only yellow markings, there would be a chance that the internal repetitions were actually only in the selected clip. Therefore the repetition information gathered from the data would be redundant, as the test data (current clip) is never included with the training data (the current translation memory in this context), and there would be no match saved in the training corpus (in terms of training the EBMT system and recalling saved translations). Therefore the EBMT system would be unable to provide a previously saved translation. If, however, the clip chosen had green markings, it would not matter if the repetition occurred in the clip itself, as the green marking indicates that there are more occurrences of this segment elsewhere in the other corpora, and hence in the training data, allowing for the repetition to be found by the EBMT system. If a segment is colour-coded red (external repetition/match), it means it does not re-appear in the first *Harry Potter* film, and therefore it would never be the case that the same red segment would appear twice in the same clip.

**Figure 4.1:** Examples of colour-coded repetitions found in the corpora. Uncoloured SL segments were not repeated in any of the corpora.

| English Subtitles | German Subtitles |
|---|---|
| All you do is walk straight at the wall between platforms 9 and 10. | Du musst geradewegs auf die Wand zulaufen. Zwischen Gleis 9 und 10. |
| Alohomora! | Alohomora! |
| Alohomora. | Alohomora. |
| Alohomora. | Alohomora. |
| Alohomora? | Alohomora? |
| Also, our caretaker, Mr. Filch, has asked me to remind you | Des Weiteren bat mich Mr. Filch, unser Hausmeister, euch an eins zu erinnern. |
| and a stranger just happens to have one. | und ein Fremder taucht auf und hat zufällig einen dabei. |
| Apparently not. | Offensichtlich ein Irrtum. |
| Apparently not. | Sieht nicht so aus. |
| And a thirst to prove yourself. | Und den Drang, sich zu beweisen. |
| and agreed it was best all around. | und waren uns einig, das sei das Beste. |
| And between you and me, that is saying something. | Und mal unter uns beiden: Das will schon etwas heißen. |
| And does Mr. Harry Potter have his key. | Hat denn Mr. Harry Potter auch seinen Schlüssel dabei. |
| and even put a stopper in death. | sogar den Tod verkorkt. |
| and exact art that is potion making. | und exakte Kunst der Zaubertrankbrauerei. |
| And for good reason. | Und zwar aus gutem Grund. |
| And I have a few last minute points to award. | Daher habe ich noch ein paar letzte Punkte zu vergeben. |

Based on this information we selected the ten most 'colourful' clips which provided us with various examples of internal and external repetitions (or 100% matches). We then calculated internal and external repetitions for each of the ten clips, with the intention of choosing the 'best' six. For the ten movie clips internal repetitions were calculated by comparing each clip with an empty TM, firstly at segment level (see Table 4.2 below), and secondly at sub-segment level (see Table 4.3 below). 100% matches are calculated by comparing each input clip with three different TMs, slightly modified versions of Corpus A, B and C: Corpus A minus the *Harry Potter* input clip (**Corpus AM**), Corpus B minus the *Harry Potter* input clip (**Corpus BM**), and Corpus C minus the *Harry Potter* input clip (**Corpus CM**). These results are presented below in Tables 4.4 (segment-level matches) and 4.5 (sub-segment level matches). These tables show the number of repetitions contributed by the individual corpora (HPCM, LOTRC and

MGC) and the combined number of repetitions (Corpus AM, Corpus BM and Corpus CM).

In addition to segment-level matches, we also calculated internal repetitions and 100% matches at sub-segment level (see Table 4.3). We manually segmented the subtitles in each of the ten movie clips based on the Marker Hypothesis approach to segmentation (sub-segments). An increase in sub-segment level repetitions and 100% matches between the clips and the corpora also indicates the potential leverage we might expect from an EBMT system using these particular source texts (clips) as our test data and the different corpora as our training data.

**Table 4.2:** Number of internal **segment level** repetitions per movie clip

| Clip number | Internal repetitions |
|---|---|
| 1 | 0 |
| 2 | 3 |
| 3 | 0 |
| 4 | 1 |
| 5 | 4 |
| 6 | 0 |
| 7 | 0 |
| 8 | 0 |
| 9 | 1 |
| 10 | 0 |

**Table 4.3:** Number of **sub-segment level** internal repetitions per movie clip (based on the Marker Hypothesis)

| Clip number | Internal repetitions |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 0 |
| 4 | 5 |
| 5 | 6 |
| 6 | 1 |
| 7 | 7 |
| 8 | 2 |
| 9 | 5 |
| 10 | 0 |

**Table 4.4:** Number of 100% **segment level** matches between each movie clip and the three corpora

| **Clip number** | HPCM | **Corpus AM** HPCM | LOTRC | **Corpus BM** HPCM +LOTRC | MGC | **Corpus CM** HPCM + LOTRC + MGC |
|---|---|---|---|---|---|---|
| 1 | 6 | **6** | 2 | **8** | 5 | **13** |
| 2 | 8 | **8** | 5 | **13** | 7 | **20** |
| 3 | 8 | **8** | 2 | **10** | 6 | **16** |
| 4 | 7 | **7** | 3 | **10** | 7 | **17** |
| 5 | 15 | **15** | 5 | **20** | 9 | **29** |
| 6 | 5 | **5** | 2 | **7** | 3 | **10** |
| 7 | 9 | **9** | 2 | **11** | 8 | **19** |
| 8 | 3 | **3** | 1 | **4** | 2 | **6** |
| 9 | 7 | **7** | 6 | **13** | 4 | **17** |
| 10 | 5 | **5** | 2 | **7** | 5 | **12** |
| **Total** | 73 | **73** | 30 | **103** | 56 | **159** |

**Table 4.5:** Number of 100% **sub-segment level** matches between each movie clip and the three corpora (based on the Marker Hypothesis)

| **Clip** | HPCM | **Corpus AM** HPCM | LOTRC | **Corpus BM** HPCM +LOTRC | MGC | **Corpus CM** HPCM + LOTRC + MGC |
|---|---|---|---|---|---|---|
| **1** | 326 | **326** | 46 | **372** | 374 | **746** |
| **2** | 261 | **261** | 13 | **274** | 750 | **1024** |
| **3** | 150 | **150** | 27 | **177** | 607 | **784** |
| **4** | 256 | **256** | 21 | **277** | 174 | **451** |
| **5** | 185 | **185** | 32 | **217** | 321 | **538** |
| **6** | 78 | **78** | 15 | **93** | 403 | **496** |
| **7** | 310 | **310** | 181 | **491** | 1732 | **2223** |
| **8** | 329 | **329** | 195 | **524** | 597 | **1121** |
| **9** | 381 | **381** | 263 | **644** | 988 | **1632** |
| **10** | 190 | **190** | 120 | **310** | 657 | **967** |
| **Total** | 2466 | **2466** | 913 | **3379** | 6603 | **9982** |

We can see from Tables 4.2 and 4.3 that some of the movie clips contain no internal segment or sub-segment repetitions. Table 4.3 shows some clips that contain no segment repetitions, but they contain sub-segment repetitions. Of the clips that contain segment repetitions, the corresponding number of sub-segment repetitions either stays the same or increases. Tables 4.4 and 4.5 indicate a high number of external repetitions for all clips, at segment level on the one hand, and at sub-segment level on the other. We carried out this analysis to investigate whether we could detect a higher number of segment and sub-segment repetitions and matches if we increased the corpus size, while systematically introducing subtitles from non-subject specific genres. In all cases increasing the training data and simultaneously decreasing the homogeneity of the corpus brought about an increase in the number of SL repetitions. Nonetheless, the most 'homogeneous' corpus (HPCM) offered the highest number of SL repetitions and therefore accounted for the highest proportion of repetitions when the three corpora were combined. Table 4.6 below shows the percentage increase in SL repetitions (segment-level) from corpus AM to corpus BM and from corpus AM to corpus CM. It also shows the percentage increase in corpus size for the same pairings. These results show that even though there is a clear increase in the number of SL repetitions when we increase the training data and simultaneously decrease the homogeneity of the corpus, the increase of SL repetitions is not commensurate with the increase in corpus size.

**Table 4.6:** Percentage increases in repetitions (segment-level) and corpus size between Corpus AM and the other two corpora, Corpus BM and Corpus CM

| | **CORPUS BM** | **CORPUS CM** |
|---|---|---|
| % increase in repetitions from Corpus AM | **41%** | **118%** |
| % increase in corpus size from Corpus AM | **63%** | **604%** |

These quantitative data are helpful for comparison purposes. They say nothing, however about the linguistic content of the corpus, and the increase in repetitions may have no bearing on the kind of quality we require to increase the intelligibility and/or acceptability levels of the subtitles.

Therefore, the next section describes the qualitative study we carried out to take this analysis a step further. We located each of the repetitions from the three corpora, and presented these subtitles along with their TL translations to three subjects, who subjectively rated the reusability of the TL translations in the context of *Harry Potter*.

## 4.2 Human Evaluation of the Reusability of TL Subtitles

The next few sections present the data gathered during the sessions with three subjects to evaluate the reusability of target language subtitles. Firstly, we discuss the Clips Booklets data. When analysing the data from the Clips Booklets we are interested in the number of EBMT subtitles deemed acceptable by the human judges in the context given. As already indicated, this helps us chose the six movie clips for the retrospective phase, as we want to choose six clips that have a high number of SL repetitions and contain a high number of acceptable TL translations that can be used in different contexts (as indicated by our three subjects). In addition we want to know which of the three corpora generated the EBMT subtitles that were deemed most acceptable by the subjects.

We begin by looking at the data generated by the Clips Booklets. These booklets contain 61 sets[84] (183 subtitles) of EBMT subtitles. For each subtitle, subjects were asked to select whether the subtitle was acceptable in the context given. If they were unsure, they could tick the 'don't know' option.

Table 4.7 below shows the distribution of responses across the three corpora per EBMT-generated subtitle. It shows that the subtitles generated using Corpus BM received the highest number of 'yes'[85] responses and the lowest number of 'no' responses in relation to the acceptability of the subtitles in the particular context. The subtitles generated by this corpus also received a slightly higher number of 'don't know' responses.

[84] A set consists of three subtitles representing the three corpora used to generate the subtitles.

[85] We need to mention here that in many of the cases in the Clips Booklets, the same translation was chosen by the EBMT system from all three corpora. Therefore, in the cases where the subjects all gave a 'yes' or 'no' or 'don't' know' response for all three translated subtitles chosen by the EBMT system, they were, in fact, approving or disapproving of the same translation three times. However, the sets in the Options Booklets never contained the same translations, as these were alternative translations located in the corpora.

**Table 4.7**: Overall subject responses for Clips Booklets in relation to the acceptability of the TL translations generated by the EBMT system

| **Acceptable EBMT-Subtitle** | **Corpus AM** HPCM | **Corpus BM** HPCM + LOTRC | **Corpus CM** HPCM + LOTRC +MGC |
|---|---|---|---|
| Yes | 120 | 126 | 119 |
| No | 54 | 46 | 55 |
| Don't know | 9 | 11 | 9 |
| **Total** | **183** | **183** | **183** |

It was noted from the Clips Booklets that 51 Corpus BM translations (83%) from a possible 61 were the same as Corpus AM translations; 41 Corpus CM translations (67%) from a possible 61 were the same as Corpus AM translations. These figures suggest a strong contribution of Corpus AM data in the translations generated by the two other corpora. These subtitles could be exact matches between the test and training data, or they could be a result of input subtitles that were broken into sub-segments, matched with sub-segments in the training corpora and then recombined. Either way this result emphasises the importance of subject-specific data (strong homogeneity) in the acceptance of EBMT-generated subtitles (in this phase).

From the corpus data presented above, we saw that Corpus CM exhibited the highest number of SL repetitions. However, Corpus BM was deemed to have produced more acceptable subtitles in this context. A detailed breakdown of how the responses from these sessions were distributed across each movie clip is included in Appendix H. Here we present some observations in relation to the ten Clips Booklets:

- In 47.5% of the cases there is agreement among all three evaluators:
  - In 39.3% of cases (24 sets) all subjects agreed that all the translations offered were acceptable in the given context ('yes' response)
  - In 8.2% of cases (5 sets) all subjects agreed that all the translations offered were unacceptable in the given context ('no' response)
- In 13.1% of cases (8 sets) all subjects gave a yes response to two of the three translations offered

- ❖ In 27.8% of cases (17 sets) two subjects agreed on the same acceptability response for all three translations:
  - 16.3% of cases (10 sets) two subjects agreed on a yes response
  - 11.4% of cases (7 sets) two subjects agreed on a no response

From these results we can see that in 93.5% of the cases (or 56 sets) at least two subjects agreed on the same response. In 68.8% of cases (or 42 sets) at least two of the three subjects considered two of the three translations acceptable in the given context. These results show strong inter-subject agreement. Table 4.8 shows the number of acceptable responses per corpus for each subject.

**Table 4.8:** The number of subtitles per corpus that are deemed acceptable by each subject

| **'Yes' responses** | **Corpus AM** HPCM | **Corpus BM** HPCM + LOTRC | **Corpus CM** HPCM + LOTRC +MGC |
|---|---|---|---|
| **Subject 1** | 43 | 45 | 44 |
| **Subject 2** | 38 | 40 | 36 |
| **Subject 3** | 39 | 41 | 39 |
| **Total 'Yes' responses** | **120** | **126** | **119** |

Using these acceptable response scores, we can now rank the ten clips based on the two selection criteria outlined earlier, namely the highest number of acceptable EBMT-translated subtitles and the number of different EBMT-generated subtitles per clip. Table 4.9 shows the ten clips ranked in order. The top six clips were later used in the retrospective evaluation sessions.

**Table 4.9**: Clips ranked according to the acceptable responses and number of different subtitles

| Rank | Clip | Acceptable Responses | Number of different EBMT-generated subtitles |
|---|---|---|---|
| 1 | **5** | **71** | **9** |
| 2 | **3** | **55** | **9** |
| 3 | **7** | **55** | **8** |
| 4 | **9** | **37** | **6** |
| 5 | **4** | **33** | **5** |
| 6 | **10** | **32** | **4** |
| 7 | 1 | 30 | 6 |
| 8 | 2 | 45 | 3 |
| 9 | 6 | 15 | 3 |
| 10 | 8 | 4 | 2 |

Table 4.10 below presents the clips chosen for the next phase showing levels of repetition between subtitles in the clips and the three corpora.

**Table 4.10**: The six clips chosen for the retrospective phase and their corresponding levels of repetition[86] vis-à-vis the three training corpora

| **Clip number** | HPCM | **Corpus AM** (HPCM) | LOTRC | **Corpus BM** (HPCM +LOTRC) | MGC | **Corpus CM** (HPCM + LOTRC + MGC) |
|---|---|---|---|---|---|---|
| 5 | 15 | **15** | 5 | **20** | 9 | **29** |
| 7 | 9 | **9** | 2 | **11** | 8 | **19** |
| 4 | 7 | **7** | 3 | **10** | 7 | **17** |
| 9 | 7 | **7** | 6 | **13** | 4 | **17** |
| 3 | 8 | **8** | 2 | **10** | 6 | **16** |
| 10 | 5 | **5** | 2 | **7** | 5 | **12** |
| **Total** | 51 | **51** | 20 | **71** | 39 | **110** |

[86] Repetition here means 'match' with the corpus, i.e. only external repetitions.

Table 4.10 shows that in almost all cases where there were high levels of repetition in a clip with respect to the corpora, there were high levels of subject agreement on the acceptability of the EBMT-generated subtitles for these clips. However, in some cases the 'most acceptable' clips did not contain the highest number of repetitions.

Before we look at the results from the Options Booklets we will comment briefly on the characteristics of the repeated subtitles presented in the 'best six' clips in the Clips Booklets. The repeated subtitles were on average 2.23 words in length, compared to an average subtitle length of 4.92 for the entire movie. This might suggest that the repeated subtitles were short phrases.

We will now move on to discuss the results from the Options Booklets. As with the Clips Booklets, when we analysed the data from the Options Booklets we were interested in the number of 'acceptable' responses, indicating that a translation located in the corpora could be used in a different context to the one it had originally been used in (for example a subtitle from the romantic comedy *As Good as it Gets* could be used in the context of the *Harry Potter* movie clip given in the corresponding Clips Booklets). As before with the Clips Booklets, we were also interested in the corpus in which these 'acceptable' translations were located. As indicated in section 3.4.4, the Options Booklets contained 40 sets of subtitles, with each set containing between two and twelve alternative subtitle translations extracted from the corpora, but which had not been chosen by the EBMT system. Table 4.11 shows the number of alternative translations offered by Corpora AM, BM and CM, and the contribution of each "additional" sub-corpus (in parentheses). From the data in Table 4.11 we can see that Corpus AM offered the highest proportion (relative to corpus size) of alternative translations.

**Table 4.11**: Number of alternative translations offered by the complete corpora with the contribution of each "additional" sub-corpus (in parentheses)

| **Corpus AM** (HPCM) | **Corpus BM** HPCM + (LOTRC) | **Corpus CM** HPCM + LOTRC + (MGC) |
|---|---|---|
| 58<br>(HPCM 58;<br>8.52 per 1,000 words) | 60<br>(LOTR 2;<br>0.180 per 1,000 words) | 123<br>(MGC 63;<br>1.31 per 1,000 words) |

Table 4.12 below shows the number of alternative translations accepted by each of the three subjects out of the possible total number presented in the Options Booklets (given in Table 4.11). The columns in bold correspond to the complete corpora and show the number of alternative translations accepted by each subject and the percentage these translations make up of the total number of alternative translations (as indicated in Table 4.11). These columns also indicate the number of the alternative translations relative to corpus size (per 1,000 words) (see Appendix I for a breakdown of subject judgements per movie clip). Columns that are not highlighted in bold show the contribution of HPCM to the AM score, LOTRC to the BM score, and MGC to the CM score.

**Table 4.12**: Accepted translations from the Options Booklets (per corpus and subject) that can be used in the given context of the *Harry Potter* subtitle

| **Subject** | HPCM | **Corpus AM** | LOTRC | **Corpus BM** | MGC | **Corpus CM** |
|---|---|---|---|---|---|---|
| **1** | 21 (36.2%) | **21 (36.2%) (3.08 per 1,000 words)** | 1 (50%) | **22 (36.6%) (1.98 per 1,000 words)** | 29 (46%) | **51 (41%) (1.06 per 1,000 words)** |
| **2** | 23 (39.6%) | **23 (39.6%) (3.38 per 1,000 words)** | 1 (50%) | **24 (40%) (2.16 per 1,000 words)** | 27 (42.8%) | **51 (41%) (1.06 per 1,000 words)** |
| **3** | 30 (51.7%) | **30 (51.7%) (4.41 per 1,000 words)** | 1 (50%) | **31 (52%) (2.79 per 1,000 words)** | 29 (46%) | **60 (49%) (1.25 per 1,000 words)** |
| **Totals** | 73 | **73** | 3 | **77** | 85 | **162** |

From the data in Table 4.12 we can see that while Corpus CM contributes the highest number of alternative translations deemed acceptable, Corpus AM contributes most acceptable alternative translations relative to corpus size (shown by the 'per 1,000 words' figure). We note that subjects 1 and 2 accepted more alternative translations contributed by MGC than by HPCM; in contrast subject 3 accepted one more alternative translation offered by HPCM than offered by MGC.[87]

[87] In relation to the acceptable percentage each of the subcorpora contributes (HPCM, LOTRC, and MGC respectively) we can see the highest percentage contribution is from the LOTRC. The subjects accepted 50% of the LOTRC contributed translations. That said this figure has to be ignored due to the extremely low number (2) of alternative translations offered by the corpus.

The results given in Table 4.12 shed some light on the data in Table 4.7. Corpus BM has a weaker homogeneity, is larger in size and contains more SL repetitions than Corpus AM. Corpus CM is even larger in size and contains a higher number of SL repetitions, and has a weaker homogeneity than Corpus BM. That said Corpus CM was not deemed to have produced the most acceptable EBMT subtitles (albeit by very small amounts), but it was the corpus that was deemed to have contained the highest number of alternative translations acceptable for use in the *Harry Potter* context. This result shows the value of combining subject-specific data with less homogeneous data, increasing corpus size and by doing so increasing the level of SL repetitions which resulted in a corpus containing the highest number of reusable TL translations in new contexts. If we hypothesise that repetition and perceived reusability are two factors of a corpus that contribute to increasing acceptability of MT output, Corpus CM is deemed to be the corpus that has the most potential. However, the fact that Corpus BM was judged to have generated the most 'acceptable' EBMT subtitles casts some doubt on this hypothesis.

The results also suggest that the homogeneity of the corpus affects the acceptability of the subtitles in a given context, which has been shown by Armstrong (2007). However, in his study homogeneity was compared between subtitle text and non-subtitle text (with subtitle text being more useful to train the EBMT system if generating new subtitles). The current study, on the other hand, deals only with subtitle text and the results suggest that adding increasing amounts of non-genre specific data to increase the corpus size improves the possibility of obtaining acceptable subtitles in a particular context. This is a preliminary observation from the prospective phase (in a non-AVT environment), and we test its accuracy during the retrospective evaluation sessions (in an AVT environment).

## 4.3 Summarising Results

The data collected during the prospective phase were used to create corpus profiles and to enable investigation of the relationship between the corpora used to train the EBMT system and the intelligibility and acceptability of EBMT subtitles. We observed the following results:

- An increase in corpus size resulted in an increase in SL segment and sub-segment repetitions. As these two variables were increased, the homogeneity of the corpus was decreased
- In addition there was an increase in the number of alternative TL translation segments in the corpus deemed acceptable in the given (*Harry Potter*) context (e.g. the highest number of alternative TL translation segments were contained in Corpus CM)
- That said repeated segments and alternative translations do not increase in direct proportion to corpus size, with Corpus AM offering the highest proportion of alternative translations relative to corpus size
- However, Corpus BM was deemed to have generated the most 'acceptable' EBMT subtitles. Even though Corpus CM contained the highest number of alternative translations judged acceptable by the human evaluators, the EBMT system did not make full use of these repeated examples. Thus Corpus BM was deemed to have generated the most acceptable subtitles
- We must also note that 83% of Corpus BM (German) subtitles were the same as Corpus AM subtitles (possible exact matches or recombined sub-segments), which highlights the importance Corpus AM played in the overall acceptability judgements
- Considering the contribution from each of the 'sub-corpora', the number of alternative translations relative to the size of Corpus BM and Corpus CM represents low 'added value'

The prospective phase results show that increasing the corpus size and the number of SL repetitions, while decreasing the homogeneity do not result in increased acceptability of machine-generated subtitles. Given that these evaluations were conducted in a non-AVT environment, we need to conduct the retrospective phase before finalising our conclusions.

## 4.4 Concluding Remarks

This chapter presented the results from the prospective phase of the study. There were two stages in this phase: the first stage was an analysis of the corpora, and we generated corpus profiles of the corpora used to train the EBMT system. The second stage

gathered human judgements on the reusability of TL segments in the *Harry Potter* context. The prospective phase allowed us to conduct corpus profiling, and from this to make claims about the acceptability of machine-generated subtitles before conducting the end-user evaluation. In the next chapter we analyse the results from the retrospective phase, and test any claims made during this phase.

# - Chapter 5
# - Retrospective Phase: Results and Analysis

# 5 Retrospective Phase: Results and Analysis

This chapter presents the results from the retrospective phase and is divided into two sections, the first section (5.1) presenting the quantitative analysis results. For this analysis we ran statistical tests within the four quality characteristic categories defined in Chapter 3 (comprehensibility, readability, style and well-formedness), looking at significant inter-corpus differences. This is followed by an intra-corpus analysis, taking soundtrack language, prior knowledge (PK) and linguistic background (LB) into consideration. We then present a qualitative analysis (5.2), where we once again analyse the data within the four quality characteristic categories plus an additional *overall* category. When working with the qualitative data, additional themes emerged from the data set and we discuss these in relation to the intelligibility and acceptability of EBMT subtitles. Before we summarise our results, we provide automatic metric scores for our data and make comparisons with previous MT subtitle evaluation studies. The chapter concludes with a summary of the results and a discussion of our findings.

## 5.1 Quantitative Analysis

During the retrospective evaluation sessions we collected quantitative and qualitative data using the interview questionnaire approach, as outlined in Chapter 3. Firstly this chapter examines the responses given by the subjects from a quantitative viewpoint, using statistical techniques to analyse the data. The software we use to conduct the statistical tests is SPSS (Statistical Package for the Social Sciences), one of the standard statistical analysis applications used within the academic community (Moore 2006).

When examining our quantitative data we use descriptive statistics, which allow us to describe the characteristics of the sample in the study, to check the variables for any violation of the assumptions underlying the statistical techniques used to address the research questions, and to address specific research questions. In addition we use statistical techniques to compare groups (Pallant 2005:49). An alpha of .05 or less is adopted for all statistical tests.

The data (quantitative and qualitative) collected during the evaluation sessions are used to answer the research questions posed earlier:

- From an end-users' perspective:

  Does

  1. increasing levels of SL repetitions between the test and training data,
  2. increasing the size of the corpus, and
  3. decreasing the homogeneity of the corpus

have a significant impact on the intelligibility and acceptability (as operationalised in our four quality characteristics) of EBMT-generated subtitles?

The three independent variables (repetition levels, size, homogeneity) are manifest in each of the corpora, and therefore we compare inter-corpus results for each of the four quality characteristics.

### 5.1.1 Subjects

We described earlier in section 3.4.2 how we recruited subjects for the retrospective phase. Here we mention briefly the breakdown of subjects in terms of gender, age, educational background and linguistic background (Figure 5.1).

**Figure 5.1:** Characteristics of subjects who participated in retrospective evaluation sessions

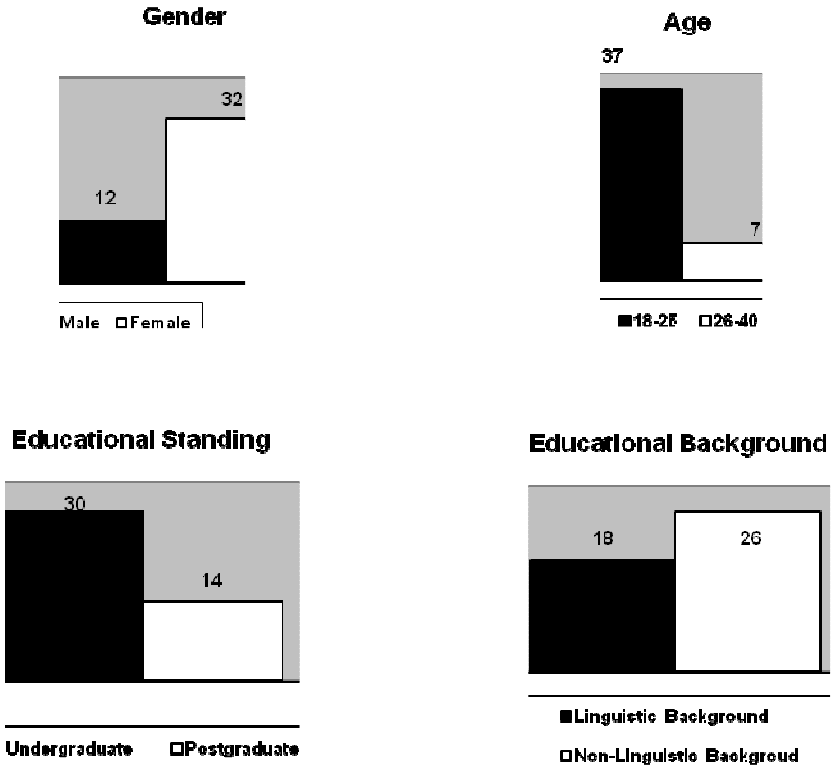



Figure 5.1 presents a breakdown of subjects who participated in the retrospective end-user evaluation sessions. The preponderance of female students and students in the 18-24 age bracket is explained by well-known trends in registration for courses in humanities and social sciences (UNESCO Institute for Statistics, European Commission), from which DCU's German-speaking exchange students are mostly drawn,[88] and by the fact that exchange students are predominantly undergraduates. Subjects with a non-linguistic background are mainly business students and engineering/science students. Students with a linguistic background are mainly translation studies students.

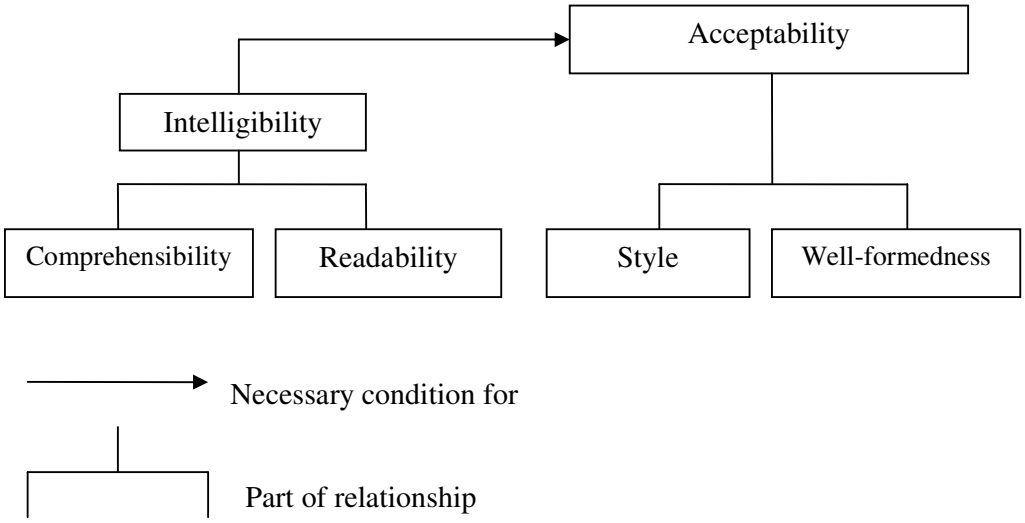
### 5.1.2 Data Collection

As argued in Chapter 2, intelligibility is a necessary but not a sufficient condition for acceptability of subtitles. To measure intelligibility, we measure the comprehensibility

[88] In addition, a lower number of women pursue courses in fields such as engineering (Womeng 2005), but research in the US has shown that of the engineering-based courses, even though a lower percentage of women actually study in this area (National Science Foundation 2008), a higher percentage will opt for the study abroad programme (Global Engineering Education Exchange 2009).

and readability of the subtitles. In addition, to evaluate acceptability we also measure their style and well-formedness. The relationship between these variables is shown in Figure 5.2 below.

**Figure 5.2:** A breakdown of the characteristics we measure to establish the intelligibility and acceptability of the EBMT subtitles




There are two types of variables in the data set for the quantitative study: categorical and continuous. When deciding on statistical tests to run it is important to check if the test is appropriate for the type of variable being examined. The four continuous variables in the study are: comprehensibility, style, errors (investigated under well-formedness) and satisfaction (investigated under *overall* category).[89] For each of these variables an interval scale was used to collect the data. All other variables in the data are coded as categorical.

### 5.1.3 Data Preparation

Before conducting any statistical analysis we generated descriptive statistics on the variables to ensure the data were not violating any of the assumptions made by the

[89] The overall category mentioned in Chapter 3 includes questions that relate to all six movie clips considered together. We asked the subjects how satisfied they were with the subtitles, taking all clips into account.

individual tests. We ran a Frequencies test on the categorical variables to check for values that fall outside the possible ranges. Given that the researcher filled in the data set from interview questionnaires, there were no missing data, and the descriptive statistics showed no data had been wrongly entered. As the statistical techniques we used in this study assume that the distribution of scores on the dependent variables is 'normal', we assessed the normality of our data before proceeding. Using the Descriptives test, we generated 'summary' statistics (e.g. mean, median, standard deviation) for the continuous variables to assess the normality of the data and to check for outliers (cases with values well above or well below the majority of other cases (Pallant 2005:58)). Some outliers were present in the data, but as the mean and 5% trimmed mean values for each variable are similar (as shown in Table 5.1), this indicated that the outliers were not having a lot of influence on the mean score, and therefore we kept these values in our data set (ibid). From the results the distribution of scores for our dependent variables was reasonably 'normal'.

**Table 5.1**: Mean and 5% trimmed mean for continuous variables

| **Continuous Variable** | **Mean** | **5% Trimmed Mean** |
|---|---|---|
| Comprehensibility | 3.2121 | 3.2221 |
| Errors | 2.7803 | 2.7475 |
| Style | 3.6174 | 3.6818 |
| Satisfaction (overall) | 2.5795 | 2.5884 |

The groups of subjects per corpus were almost the same, which is important when conducting statistical tests. For other variables such as having a linguistic background, the groups were not as evenly divided, but this was taken into account by the statistical test. Another factor that needed consideration was the 'independence of observations', meaning that observations or measurements must not be influenced by any other observation or measurement. As we conducted all of the interview questionnaire-type sessions on an individual basis, this assumption of independence was not violated.

In the next section we outline the research questions we wanted to address and the statistical techniques used to analyse them.

### 5.1.4 Statistical tests to compare groups

To test the relationship between the independent variables and the dependent variables, we ran tests between the corpora and the numerous variables that combine to measure intelligibility and acceptability. Table 5.2 overleaf (continued on pg 180) outlines the variable (within the four characteristic groups), the type of variable, the type of statistical technique used, and whether or not the result was statistically significant. If a result was significant, the *p*-value is given in each case and if the result was not significant, this is indicated by *n/s*. The data set includes responses from 44 subjects, each of whom viewed six clips, meaning there were 264 responses in total (44 x 6).

**Table 5.2**: Variables used for the statistical analyses, the statistical techniques and the results of the statistical tests

| **Comprehensibility** | | | |
|---|---|---|---|
| **Questions** | **Variable Type** | **Statistical Test** | **Result** |
| Did the subjects understand the clip? | Categorical | Chi-Square test for independence | n/s |
| Did the subjects answer the internal questions correctly? | Categorical | Chi-Square test for independence | n/s |
| Did the subjects use the subtitles to understand the clip? | Categorical | Chi-Square test for independence | n/s |
| Comprehensibility Scale | Continuous | One-way ANOVA | n/s |
| **Readability** | | | |
| **Questions** | **Variable Type** | **Statistical Test** | **Result** |
| Was the speed of the subtitles suitable? | Categorical | Chi-Square test for independence | p=.004<br><br>27.8% of Corpus BM subjects thought the speed of the subtitles was unsuitable, which is significantly different from the percentage of Corpus AM subjects (18.9%) and Corpus CM subjects (8.3%). |
| Did the subjects note any subtitles that seemed out of context for that clip? | Categorical | Chi-Square test for independence | n/s |

| Style | | | |
|---|---|---|---|
| **Questions** | **Variable Type** | **Statistical Test** | **Result** |
| Did anything in the subtitles particularly bother the subjects? | Categorical | Chi-Square test for independence | n/s |
| Did anything in the subtitles particularly amuse the subjects? | Categorical | Chi-Square test for independence | p=.011<br><br>The percentage of subjects (26.7%) who noted something amusing about the subtitles in Corpus AM is significantly higher than that for Corpus BM (12.2%) and Corpus CM (11.9%). |
| Style Scale | Continuous | One-way ANOVA | p=.007.<br><br>There is a significant difference between the style scale scores for corpus AM and Corpus BM. Corpus CM does not differ significantly from either AM or BM. |
| **Well-Formedness** | | | |
| **Questions** | **Variable Type** | **Statistical Test** | **Result** |
| Were the subtitles acceptable for people who do not understand the soundtrack? | Categorical | Chi-Square test for independence | n/s |
| Did the subjects note any particularly well-translated subtitles? | Categorical | Chi-Square test for independence | n/s |
| Did the subjects note any errors in the subtitles? | Categorical | Chi-Square test for independence | n/s |
| What types of errors were noted by the subjects? | Categorical | Chi-Square test for independence | p=.002 for C1 errors.<br><br>The percentage of C1 errors (28.6%) noted by subjects viewing subtitles from Corpus CM is significantly higher than the percentage noted by subjects viewing subtitles from AM (7.8%) and/or BM (11.1%). |
| Error Scale | Continuous | One-way ANOVA | n/s |

From Table 5.2 we can see that there are two different statistical techniques used to analyse the data, in this instance depending on the type of variable used. The test used to measure the relationship between a categorical variable (corpus) and a continuous variable (scale score) is the one-way analysis of variance (one-way ANOVA); the test used to measure the relationship between two categorical variables (corpus and quality characteristics) is the Chi-Square test for independence. These statistical tests are discussed in more detail in the next section.

### Parametric versus non-parametric statistics

According to Pallant (2005:286) parametric tests "make assumptions about the population that the sample has been drawn from", while non-parametric techniques "do not have such stringent requirements (as parametric tests), and do not make assumptions about the underlying population distribution." Parametric techniques are considered to be more powerful than non-parametric techniques in detecting differences between groups. In our data set we have four continuous variables. We want to compare the mean scores on these continuous variables for three different groups of subjects (three corpora). In this instance we use the one-way analysis of variance (one-way ANOVA) technique. This technique "compares the variance (variability in scores) *between* the different groups (believed to be due to the independent variable) and the variability *within* each of the groups (believed to be due to chance)" (ibid:214). This technique allows us to statistically test a question such as: Are subjects who view subtitles generated by Corpus AM more likely to rate them as more comprehensible than subjects viewing subtitles generated by either Corpus BM or Corpus CM?

However, there are circumstances when non-parametric techniques are the most suitable, for example if you are working with data measured on nominal (categorical) and ordinal (ranked) scales or if you are working with small samples. There is a wide variety of non-parametric techniques for a number of situations. The technique which is most appropriate to use with our data is the Chi-Square test for independence. This test is used to determine whether two categorical variables are related, by "comparing the frequency of cases found in the various categories of one variable across the different categories of another variable" (ibid). In the current study the corpora are coded as categorical variables (1, 2 or 3), and also the questions relating to *readability*. In the other three quality characteristic groups (*comprehensibility*, *style* and *well-formedness*),

all but one of the questions are coded as categorical variables. For the question regarding the types of errors noted by the subjects (*well-formedness*), we adopted Flanagan's (1994) MT error classification, and grouped the errors noted by subjects into three classes: Class 3 (C3) contains the errors that most affect the intelligibility of subtitles and Class 1 (C1) contains errors that least affect the intelligibility. We matched each error noted by the subjects to a class and listed the total number of errors by class within each corpus (see Table 5.3 below for details on error classification).

**Table 5.3:** Flanagan's (1994) MT Error Classification Table

| **Class 1 (least effect)** | **Class 2** | **Class 3 (most effect)** |
|---|---|---|
| **accent** | **not-found-word** | **expression** |
| **capitalization** | **verb inflection** | **category** |
| **spelling** | **noun inflection** | **rearrangement** |
| **article** | **other inflection** | **word selection** |
| | **pronoun** | **conjunction** |
| | **preposition** | |
| | **agreement** | |
| | **negative clause** | |
| | **boundary** | |

The Chi-Square test for independence technique allows us to statistically test a question like: Did subjects who viewed Corpus AM subtitles consider the speed of the subtitles more suitable than subjects who viewed either Corpus BM or Corpus CM subtitles?

### 5.1.5 Analysis of Responses

We will now look at the four quality characteristics individually and comment on the results given in Table 5.2 above.[90]

#### Comprehensibility

We can see from the statistical test results outlined in Table 5.2 that none of the variables that measure *comprehensibility* differ significantly between the three corpora. Therefore based on statistical tests we cannot say subjects who viewed subtitles from one particular corpus found the subtitles more comprehensible than the subtitles from the other two corpora.

#### Readability

We can see from the statistical tests that there is one categorical variable that is significantly different within *readability*. This variable was investigated using the question:

- Was the speed of the subtitles suitable? (yes/no)

As our result is statistically significant (p=.004), this means that the proportion of Corpus BM subjects who believed the perceived speed of the subtitles was unsuitable is significantly different from the proportion of Corpus CM subjects who believed it was unsuitable. Corpus AM does not differ significantly from either of the other two. The actual speed of the subtitles was the same for each corpus. Table 5.4 shows the inter-corpus percentage breakdown of the yes/no answers relating to subtitle speed.

**Table 5.4**: The inter-corpus percentage breakdown of the yes/no answers relating to subtitle speed

| Corpus | | Subtitle Speed Suitable | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| **AM** | % within Corpus | **81.1%** | **18.9%** | **100%** |
| **BM** | % within Corpus | **72.2%** | **27.8%** | **100%** |
| **CM** | % within Corpus | **91.7%** | **8.3%** | **100%** |

[90] For each of the Tables 5.4 – 5.8, the percentages highlight the differences between the groups, as Corpus AM and BM have 15 subjects each (90 responses), while Corpus CM has 14 subjects (84 responses).

We can see that Corpus BM reported the highest percentage of answers deeming the speed of the subtitles to be unsuitable. Corpus CM shows the lowest percentage of negative responses towards the subtitle speed. From these results we could tentatively say that the readability of the subtitles generated by Corpus CM is higher than the other corpora.

### Style

From the statistical tests we see two categorical variables that are significantly different within *style*. These variables were investigated using the questions:

- Did anything in the subtitles particularly amuse the subjects? (yes/no)
- On a scale of 1-6 where did the subjects rate the style of the subtitles? (1 being very inappropriate style – 6 being very appropriate style)

Firstly we look at the question regarding the subtitles that were amusing to subjects. The result for this question is statistically significant (p=.011), meaning the proportion of subjects who noticed something amusing in the subtitles generated by Corpus AM is significantly different from the proportion of subjects from either Corpus BM or Corpus CM. Many of the responses to this question regarding amusing subtitles overlapped with responses given elsewhere in relation to linguistic issues and unusual sentence structure of the subtitles. However, some responses were related to positive aspects of the subtitles, which we discuss in more detail during the qualitative analysis. Table 5.5 shows the percentage of subjects who thought that there was something amusing about the subtitles.

**Table 5.5:** Inter-corpus percentage breakdown of subtitles deemed amusing

| Corpus | | Amusing subtitles | |
|---|---|---|---|
| | | Yes | No |
| **AM** | % within Corpus | **26.7%** | 73.3% |
| **BM** | % within Corpus | **12.2%** | 87.8% |
| **CM** | % within Corpus | **11.9%** | 88.1% |

From these results we cannot say that the style of the subtitles from Corpus AM is better or worse than the style in the other two corpora, as the subjects' comments could

represent a positive or negative attitude. We will be able to interpret these results following the qualitative analysis (see section 5.2.1).

The second question we look at is the result of the style scale. Subjects rated the style of the subtitles on a scale of 1-6, 1 being inappropriate style, 6 being very appropriate style in the given context. A one-way ANOVA test was conducted to explore the impact of the corpus on the style scores of the subtitles. We obtained a statistically significant difference (p=.007) between the style scale scores for Corpus AM and Corpus BM. Corpus CM does not differ significantly from either Corpus AM scores or Corpus BM scores. Table 5.6 below shows the inter-corpus mean scores for style, showing that Corpus BM subjects on average rate the style as more appropriate than that of Corpus AM subtitles.

**Table 5.6**: Inter-corpus mean scores for style

| Corpus | Mean | N | Std. Deviation |
|---|---|---|---|
| **AM** | **3.3556** | 90 | 1.09453 |
| **BM** | **3.8778** | 90 | .81871 |
| **CM** | 3.6190 | 84 | 1.34348 |

This finding seems slightly counterintuitive given the greater homogeneity of Corpus AM. However, we noted earlier that Corpus AM contributed the most to the EBMT-generated subtitles offered by Corpus BM and Corpus CM, and therefore factors such as prior knowledge of *Harry Potter* could have influenced this result (see Table 5.16). At this point based on the subjects' judgements on style, Corpus BM subtitles are considered to be written in the most appropriate style. This finding will be examined again during the qualitative analysis.

### Well-formedness

We can see from the statistical tests there is one categorical variable that is significantly different within *well-formedness*. This variable was investigated using the question:

- What types of errors were noted by the subjects? (C1, C2, C3)

We look at the output of the types of errors noted by the subjects. We obtained a statistically significant result (p=.002) for the number of C1 errors observed by the

subjects viewing subtitles from Corpus CM compared to the number of C1 errors noted by the subjects viewing subtitles from Corpus AM and/or Corpus BM. The number of C2 and C3 errors noted by subjects was similar across all three corpora. Table 5.7 shows the number and percentage of errors noted by subjects per corpus.

**Table 5.7**: Number and percentage of errors noted by subjects grouped by type of error

| Corpus | | C1 Errors | C2 Errors | C3 Errors | Total errors per corpus |
|---|---|---|---|---|---|
| AM | Count | 7 | 77 | 94 | 178 |
| | % within Corpus | **3.9%** | **43.2%** | **52.8%** | 100% |
| BM | Count | 12 | 79 | 85 | 176 |
| | % within Corpus | **6.8%** | **44.8%** | **48.2%** | 100% |
| CM | Count | 30 | 77 | 76 | 183 |
| | % within Corpus | **16.3%** | **42.0%** | **41.5%** | 100% |
| Total errors per type | | **49** | **223** | **255** | |

We can see from Table 5.7 that 16.3% of overall errors noted in Corpus CM were C1 errors, which is a significant increase on the percentages noted in the other two corpora (3.9% and 6.8%). There is no difference in the actual speed of the subtitles for each of the corpora, and yet it is interesting to note that the subjects who viewed Corpus CM subtitles deemed the speed of the subtitles most suitable (85.7%). These results may indicate that the subjects who viewed Corpus CM subtitles appeared to have more time to notice C1 errors present compared with the other two groups of subjects. The results also suggest something about the effect the corpus has on the number of errors spotted. We mentioned earlier that 41 of the 61 subtitles generated by Corpus CM were the same as those generated by Corpus AM. The remaining 20 subtitles could have contained many C1 errors which the subjects noticed. Nonetheless, we must also remind ourselves that the high number of C1 errors does not necessarily render subtitles from Corpus CM less 'intelligible' as these types of errors have the least effect on the intelligibility of the subtitles. This point is supported by the mean score results for errors, which did not seem to be affected by the significant difference in the number of C1 errors noted by Corpus CM, as shown in Table 5.8. Corpus AM subjects who noted the lowest number of C1 errors regarded the errors they did notice to be the most annoying on average in comparison with the other two corpora.

**Table 5.8**: Inter-corpus mean scores for errors

| Corpus | Mean | N | Std. Deviation |
|---|---|---|---|
| AM | **2.6333** | 90 | 1.18464 |
| BM | 2.8889 | 90 | 1.19403 |
| CM | 2.8214 | 84 | 1.22387 |

The results here suggest that even though the higher number of C1 errors noted by Corpus CM subjects does not necessarily render the subtitles as less intelligible, it does however suggest that these subtitles are less well-formed than the subtitles from Corpus AM or Corpus BM, thus affecting the acceptability of Corpus CM subtitles. We will address this finding when we discuss the qualitative data.

The statistical tests presented here have shown that:

- There is no difference between the subtitles in relation to comprehensibility
- The readability of Corpus CM subtitles is considered better than that of Corpus AM or Corpus BM
- The style of Corpus BM subtitles is considered the most appropriate of the three corpora
- Corpus CM subtitles are considered to be the least well-formed of the three corpora

### 5.1.6 Analysis of Additional Factors

The statistical results, presented in Table 5.2, were an analysis of responses between groups, i.e. when generating these statistics we compared groups that were looking at different sets of subtitles, depending on the corpus used (Corpus AM, Corpus BM or Corpus CM). When we conducted the evaluation sessions we had three sets of subtitles which we used as our grouping factors. In addition to this we introduced another independent variable into the three groups. We used a known language (English) and an unknown language (Dutch) for the movie soundtrack. For example, there were 15 subjects who viewed subtitles generated by Corpus AM. The order of the soundtrack

was English followed by Dutch for 8 subjects, and Dutch followed by English for 7 subjects.[91]

Therefore we conduct additional statistical tests which allow us to look at the individual and joint effect of two independent variables on one dependent variable. The first set of tests investigates the effect of corpus and language on the continuous variables: comprehensibility, errors and style[92] when the subjects view the same data set (intra-corpus investigation). This test is known as a two-way, between groups (divided by soundtrack language) analysis of variance, or a two-way ANOVA. The results for each of the variables are presented in Table 5.9 below.

**Table 5.9:** The significance, Partial Eta Squared and interaction effect results measuring the impact of corpus and language on the three continuous variables

| | **Corpus p** | **Partial Eta Squared** | **Language p** | **Partial Eta Squared** | **Corpus * Language p** |
|---|---|---|---|---|---|
| **Comprehensibility** | .289 | n/a | .040 | .017 | .727 |
| **Errors** | .385 | n/a | .021 | .021 | .767 |
| **Style** | .004 | .044 | .46 | n/a | .901 |

The column corpus * language in Table 5.9 tells us whether there is an interaction effect between the independent variables and the dependent variable, for example, that the influence of the corpus on comprehensibility scores depends on whether you are listening to a soundtrack in a known (English) or unknown (Dutch) language. The output of this effect determines how we interpret the other results. We can see from Table 5.9 that none of the scores in the corpus * language column are less than or equal to .05 meaning there is no significant difference in the effect of the corpus on any of the dependent variables for subjects listening to a known and unknown language soundtrack.

[91] When conducting statistical tests with the data from Corpus AM we had to remove data associated with one subject who was fluent in Dutch. Of the 44 subjects this was the only subject with knowledge of Dutch, other than subjects mentioning that Dutch is a similar language to German. These data were collected during the background questions section of the interview questionnaire. The data set used for these tests contains 84 responses for Corpus AM, 90 responses for Corpus BM, and 84 responses for Corpus CM.

[92] We do not include overall satisfaction here as we did not differentiate for language when asking the overall questions.

Now we check for the simple effect of one independent variable on the dependent variable.

### Comprehensibility

We can see from Table 5.9 that there is a significant main effect for language, but no significant main effect for corpus. This means that there is no significant difference in terms of comprehensibility scores within the corpora, but there is a significant difference in scores depending on which soundtrack the subject was listening to. A common way of assessing the importance of the statistically significant result from a two-way ANOVA is to calculate the effect size (Partial Eta Squared measure) which indicates the relative magnitude of the differences between means. In this case the Partial Eta Squared score (.017) given in Table 5.9 is a measurement of the effect size when using a two-way ANOVA. Cohen (1988) suggests the following guidelines when interpreting the result:

**Figure 5.3**: Cohen's guidelines for interpreting effect size

| | |
|---|---|
| .01 | small effect |
| .06 | moderate effect |
| .14 | large effect |

Using Cohen's criterion this effect size is small. Even though the effect size is small, it is worth noting that for each of the corpora, the comprehensibility of the subtitles was deemed worse when listening to the foreign language soundtrack, as presented in Table 5.10 below.

**Table 5.10:** Descriptive statistics highlighting mean scores for comprehensibility

| Corpus | Language | Mean | Std. Deviation | N |
|---|---|---|---|---|
| **AM** | Dutch | 2.9524 | .96151 | 42 |
| | English | 3.2143 | 1.25980 | 42 |
| **BM** | Dutch | 3.2667 | 1.13618 | 45 |
| | English | 3.4444 | 1.19764 | 45 |
| **CM** | Dutch | 2.9524 | 1.30575 | 42 |
| | English | 3.4048 | 1.03734 | 42 |

Table 5.10 shows that the mean scores for comprehensibility are lower in all three corpora when the subjects were listening to an unknown language soundtrack, suggesting that the known language soundtrack aided the subjects' comprehension of the movie clip, and not that the subtitles presented on a clip with an English language soundtrack were more comprehensible. The most noticeable difference between the scores is in Corpus CM, which contains approximately 77% of non-genre specific data, perhaps indicating the importance of genre-specific data when relying solely on the subtitles and image for understanding.

### Errors

In relation to errors, Table 5.9 shows a significant main effect for language, but no significant main effect for corpus. This means that there is no significant difference in terms of error scores within the corpora, but there is a significant difference in scores depending on which soundtrack the subjects were listening to. The Partial Eta Squared score (.021) suggests a small effect between the mean scores, which are presented in Table 5.11 below.

**Table 5.11:** Descriptive statistics highlighting mean scores for errors

| Corpus | Language | Mean | Std. Deviation | N |
|---|---|---|---|---|
| **AM** | Dutch | 2.5238 | 1.06469 | 42 |
| | English | 2.7619 | 1.35807 | 42 |
| **BM** | Dutch | 2.7333 | 1.00905 | 45 |
| | English | 3.0444 | 1.34765 | 45 |
| **CM** | Dutch | 2.5714 | 1.32781 | 42 |
| | English | 3.0714 | 1.06823 | 42 |

Like the scores for comprehensibility, the error scores are all lower for the clips with the unknown language soundtrack, suggesting that the errors in these subtitles are more serious than those in clips with an English language soundtrack. But this perceived severity of errors might be due to the fact that subjects do not understand the soundtrack. Once again the biggest difference between scores for known and unknown language soundtrack is observed in Corpus CM.

### Style

The results in Table 5.9 show a different story for style scores. In this case there is a significant main effect for corpus, but no significant main effect for language. This means that there was no significant difference in style scores when subjects were listening to either a known or an unknown soundtrack language, but there is a significant difference depending on the training corpus that generated the subtitles. We have already discussed this point in the previous statistical tests when we were investigating significant inter-corpus differences (e.g. significant difference between Corpus AM and Corpus BM). We should, however, point out that the scores for the subtitles with an unknown soundtrack language are all lower than the subtitles with a known soundtrack language, but these intra-corpus differences are not significant (see Table 5.12 below).

**Table 5.12:** Descriptive statistics highlighting mean scores for style

| Corpus | Language | Mean | Std. Deviation | N |
|---|---|---|---|---|
| **AM** | Dutch | 3.2857 | .99476 | 42 |
| | English | 3.3333 | 1.18253 | 42 |
| | Total | **3.3095** | 1.08635 | 84 |
| **BM** | Dutch | 3.8444 | .73718 | 45 |
| | English | 3.9111 | .90006 | 45 |
| | Total | **3.8778** | .81871 | 90 |
| **CM** | Dutch | 3.5238 | 1.36575 | 42 |
| | English | 3.7143 | 1.33043 | 42 |
| | Total | **3.6085** | 1.11840 | 84 |

The results relating to the soundtrack language imply that if subjects have knowledge of the soundtrack language they perceive the subtitles as more comprehensible (comprehensibility scale) and more well-formed (error scale). Knowing the soundtrack language does not have an impact on how the subjects measure the appropriateness of the style (although the scores were slightly higher for the English language clips). The corpus had an effect on style ratings, and Corpus AM was deemed to be the least appropriate, as discussed previously.

The next set of tests investigates the impact of language (independent variable) on the categorical questions within the four quality characteristics or dependent variables

(comprehensibility, readability, style and well-formedness). We examine whether the answers to the questions asked during the evaluation sessions (cf. Table 5.2) differ significantly depending on the soundtrack language but within the same corpus. We conduct intra-corpus tests in order to investigate the impact of the independent variable on subjects' judgements when they are viewing the same data set. We used the Chi-Square test for independence to test the impact language had on the results. After running the tests only one result was deemed statistically significant, presented in Table 5.13 below.

**Table 5.13**: Intra-corpus results deemed significantly different when differentiating for known and unknown language

| Question | | Corpus | Test | p |
|---|---|---|---|---|
| 1 | Did the subjects use the subtitles to understand the clip? (comprehension) | BM | Chi-Square test for independence | .035 |

97.8% of Corpus BM subjects said they used the subtitles to understand when listening to the Dutch soundtrack, while 82.2% of Corpus BM subjects said they used the subtitles to understand when listening to the English soundtrack.

These results suggest that Corpus BM subjects are less likely to rely on the subtitles to understand when they have knowledge of the soundtrack. This result supports the observation mentioned previously that a secondary activity such as listening to the known soundtrack is more easily achieved than a primary activity such as reading the subtitles (Koolstra et al. 2002). This result could also suggest that subjects had to use the known language soundtrack more often because the subtitles on the English language clips were less comprehensible than those on the Dutch language clips. However, as we see later, subjects were asked if they noticed a difference in the quality of subtitles depending on which soundtrack language they were listening to. 26.6% of Corpus BM viewers thought the subtitles with the English language soundtrack were of higher quality, while only 13.4% thought the subtitles with the Dutch language clips were better. This result supports Koolstra et al.'s findings.

Two other factors which might be considered as having an impact on intelligibility and acceptability of the subtitles are whether subjects have a linguistic background and/or prior knowledge[93] of the movie used in this study.

Therefore we firstly conducted statistical tests to investigate impact of corpus and linguistic background (LB) on each of the continuous variables, followed by an investigation of the impact of LB on the categorical variables. Table 5.14 presents the results of the two-way ANOVA tests for the continuous variables.

**Table 5.14:** Testing the impact of LB and corpus on the continuous dependent variables

| Continuous Variable | Statistical Test | LB p | Corpus p | Partial Eta squared | LB * Corpus p |
|---|---|---|---|---|---|
| Comprehensibility | two-way ANOVA | **.042** | n/s | **.016** | **.376** |
| Errors | two-way ANOVA | n/s | n/s | n/a | **.475** |
| Style | two-way ANOVA | n/s | n/s | n/a | **.619** |
| Satisfaction (overall) | two-way ANOVA | n/s | n/s | n/a | **.381** |

We can see that there was no interaction effect between the corpus and LB on any of the continuous variables and similarly corpus or LB did not have any effect on errors, style and satisfaction. The only significant difference recorded was for LB having an impact on comprehensibility scores (p=.042). The effect size for the comprehensibility variable (.016) can be considered small.

Table 5.15 below presents the mean scores for comprehensibility, differentiating for LB. For all three corpora, subjects with formal language training judged the subtitles to be less comprehensible than subjects with no formal language training. Corpus CM subjects with LB had the most impact on comprehensibility scores, with the average score much lower than Corpus CM subjects with no LB.

[93] In this study we consider prior knowledge to be a good understanding of the Harry Potter story and familiarity with all of the characters. This knowledge can be gained either through reading the books, watching the movies or a combination of both. Simply knowing the names of the main characters through the media is not considered as having prior knowledge.

**Table 5.15:** Mean scores for comprehensibility across corpora and differentiating for LB

| Corpus | LB | Mean | Std. Deviation | N |
|---|---|---|---|---|
| AM | Yes | **3.0741** | 1.06136 | 54 |
| | No | **3.1389** | 1.17480 | 36 |
| BM | Yes | **3.0000** | 1.27920 | 12 |
| | No | **3.4103** | 1.14456 | 78 |
| CM | Yes | **2.9048** | 1.14358 | 42 |
| | No | **3.4524** | 1.19353 | 42 |

The results presented in Table 5.15 show that subjects with formal language training deemed the subtitles to be less comprehensible. The lower scores might relate to a lower tolerance of grammatical errors when the subjects have formal training in linguistic issues.

As before, we investigate the impact of the independent variable, in this case linguistic background, on the categorical questions within the four quality characteristics. After running the tests two results were deemed statistically significant, presented in Table 5.16 below.

**Table 5.16**: Intra-corpus results deemed significantly different when differentiating for linguistic background

| Question | | Corpus | Test | p |
|---|---|---|---|---|
| 1 | Did the subjects answer the internal questions correctly? (comprehensibility) | CM | Chi-Square test for independence | .020 |
| 2 | Was the speed of the subtitles suitable? (readability) | AM | Chi-Square test for independence | .004 |

Results for the first question show that 97.2% of Corpus AM subjects with no formal language training said the speed of the subtitles was suitable, compared with 70.4% of Corpus AM subjects with formal language training, who said the speed of the subtitles was unsuitable.

Results for the second question show that 66% of Corpus CM subjects with no formal language training correctly answered the 'internal check' questions, while only 50.9% of Corpus CM subjects with formal language training correctly answered the questions.

These two significant findings show that judgements on the readability of the subtitles (Corpus AM) and the comprehensibility of the subtitles (Corpus CM) from subjects with a linguistic background did not improve the results in any way. The Corpus CM finding supports the results in Table 5.15 above.

Secondly we conducted statistical tests to investigate the impact of corpus and prior knowledge (PK) on each of the continuous variables. Table 5.17 presents the results of the two-way ANOVA tests.

**Table 5.17:** Testing the impact of corpus and PK on the continuous dependent variables

| Continuous Variable | Statistical Test | PK p | Corpus p | Partial Eta squared | PK * Corpus p | Partial Eta squared |
|---|---|---|---|---|---|---|
| Comprehensibility | two-way ANOVA | n/s | n/s | n/a | n/s | n/a |
| Errors | two-way ANOVA | n/s | n/s | n/a | **.019** | **.030** |
| Style | two-way ANOVA | n/s | **.047** | **.023** | **.026** | **.028** |
| Satisfaction (overall) | two-way ANOVA | n/s | n/s | n/a | **.019** | **.189** |

Table 5.17 shows that neither corpus nor PK had any impact on the mean scores for comprehensibility. The corpus variable had an impact on the style scores, but since there is also an interaction effect between corpus and PK on style scores, we need to firstly interpret the interaction effect. There is also an interaction effect between corpus and PK on error and satisfaction scores. In order to explore the interaction effect further we conducted an analysis of simple effects (one-way ANOVA), by looking at the results for each of the subgroups (splitting the file by PK (y/n) or by corpus (AM/BM/CM)). We looked at the effect of PK on the three dependent variables separately for each of the corpora, and the effect of corpus on the same dependent variables. Table 5.18 below outlines the significant results.

**Table 5.18:** Significant results of one-way ANOVA tests to test the interaction effect of corpus and PK on the dependent variables

| Continuous Variable | Split file by | Statistical Test | p | Eta squared[94] |
|---|---|---|---|---|
| Errors (corpus) | no PK | one-way ANOVA | .018 | 0.08 |
| Style (PK) | Corpus CM | one-way ANOVA | .028 | 0.05 |
| Style (corpus) | PK | one-way ANOVA | .001 | 0.08 |
| Satisfaction (PK) | Corpus CM | one-way ANOVA | .048 | 0.28 |
| Satisfaction (corpus) | PK | one-way ANOVA | .010 | 0.29 |

Table 5.18 shows that there are five results deemed statistically significant:

1. The mean scores for errors between Corpus AM and Corpus BM when the subjects have no PK (p=.018) (Corpus CM does not differ significantly from either of the other two; see Table 5.19).

**Table 5.19:** Inter-corpus mean scores for errors when the subjects have no PK

| Corpus | Mean | Std. Deviation | N |
|---|---|---|---|
| **AM** | **2.2917** | 1.26763 | 24 |
| **BM** | **3.2333** | 1.30472 | 30 |
| CM | 2.6111 | 1.12828 | 36 |

2. The mean scores for style between subjects who have PK and subjects who have no PK within Corpus CM (p=.028) are significantly different (see Table 5.20).

**Table 5.20:** Mean scores for style within Corpus CM, differentiating by PK

| PK | Mean | Std. Deviation | N |
|---|---|---|---|
| Yes | **3.8958** | 1.15297 | 48 |
| No | **3.2500** | 1.50000 | 36 |

[94] Like Partial Eta Squared, Eta Squared is another effect size statistic used to measure the relative magnitude of the differences between means when using a one-way ANOVA.

3. The mean scores for style between Corpus AM and the other two corpora differ significantly when the subjects have PK (p=.001) (see Table 5.21).

**Table 5.21:** Inter-corpus mean scores for style when the subjects have PK

| Corpus | Mean | Std. Deviation | N |
|---|---|---|---|
| **AM** | **3.2727** | 1.13063 | 66 |
| BM | 3.9167 | .80867 | 60 |
| CM | 3.8958 | 1.15297 | 48 |

4. The mean scores for satisfaction between Corpus AM and Corpus CM differ significantly when the subjects have PK (p=.010) (see Table 5.22).

**Table 5.22:** Inter-corpus mean scores for satisfaction when the subjects have PK

| Corpus | Mean | Std. Deviation | N |
|---|---|---|---|
| **AM** | **2.0000** | .70711 | 9 |
| BM | 2.3333 | .86603 | 9 |
| **CM** | **3.0909** | .70065 | 11 |

5. The mean scores for satisfaction between subjects who have PK and subjects who have no PK differ significantly within Corpus CM (p=.048) (see Table 5.23).

**Table 5.23:** Mean scores for overall satisfaction within Corpus CM, differentiating by PK

| PK | Mean | Std. Deviation | N |
|---|---|---|---|
| Yes | 3.0909 | .70065 | 11 |
| No | 2.0000 | 1.00000 | 3 |

The size of the effects (Cohen 1988) is medium (.08), small (.05), medium (.08), large (0.29) and large (0.28) respectively.

We will look briefly at the results where the size of the effect is considered medium or large. The first significant result suggests that subjects with no PK viewing subtitles from Corpus AM rated observed errors as more annoying than subjects with no PK viewing subtitles from Corpus BM. As these two sets of subtitles are generated by different corpora and the result refers to subjects with no PK, it could suggest the well-formedness of Corpus BM subtitles is higher than Corpus AM subtitles. We already noted that Corpus AM subjects rated observed errors as more annoying than subjects from the other corpora (see Table 5.8), and this result supports the earlier findings.[95]

Looking at the second result, subjects with PK viewing subtitles generated by Corpus AM considered the style of the subtitles less suitable than subjects with PK viewing subtitles generated by the other two corpora. We have noted previously that the style of the subtitles generated by Corpus AM was deemed the least appropriate (see Table 5.6), and considered this result somewhat counterintuitive given the input of Corpus AM subtitles in the other two groups. Investigating the effect PK has on the inter-corpus result shows that having prior knowledge of the test data (*Harry Potter* subtitles) specifically influences subjects' opinion on the appropriateness of style. The result also suggests that heterogeneity improves judgements on style for subjects who have PK, and in particular *Harry Potter*-specific expectations. The low style ratings, coupled with the earlier finding of Corpus AM subjects rating the errors as most annoying (thus reducing the well-formedness of the subtitles), suggests that Corpus AM subtitles are less acceptable than the other corpora. This finding will be examined in more detail from a qualitative perspective.

The third result looks at the inter-corpus mean scores for satisfaction when the subjects have PK. There is a significant difference between the scores given for Corpus AM subtitles and Corpus CM subtitles, with the latter rating the subtitles as more satisfactory.

The final result looks at mean scores for overall satisfaction within Corpus CM and differentiates by PK. Subjects with PK were more satisfied overall with the quality of the subtitles.

[95] It is interesting to note that of the subjects with no PK from both corpora, Corpus AM subjects noted a lower number of errors (50) than Corpus BM subjects (56), but they rated them as 'more annoying' overall.

Once again, we investigate the impact of the independent variable, in this case prior knowledge, on the categorical questions within the four quality characteristics. After running the tests three results were deemed statistically significant, presented in Table 5.24 below.

**Table 5.24:** Intra-corpus results deemed significantly different when differentiating for prior knowledge

| Question | | Corpus | Test | p |
|---|---|---|---|---|
| 1 | Did the subjects note any particularly well-translated subtitles? (well-formedness) | AM | Chi-Square test for independence | .003 |
| 2 | Did the subjects note any subtitles that seemed out of context for that clip? (readability) | AM | Chi-Square test for independence | .001 |
| 3 | Was the speed of the subtitles suitable? (readability) | AM | Chi-Square test for independence | .014 |

All of the findings given in Table 5.24 relate to the impact of prior knowledge on the categorical variables in Corpus AM.

Results for the first question show that 77.3% of subjects with prior knowledge noted well-translated subtitles compared with 41.7% of subjects with no prior knowledge.

Results for the second question show that 48.5% of subjects with prior knowledge deemed some of the subtitles to be out of context in the given movie clip, while only 8.3% of subjects with no prior knowledge deemed any of the subtitles to be out of context in the given movie clip.

Lastly, results for the third question show that only 72.4% of subjects with prior knowledge consider the (perceived) speed of the subtitles to be suitable compared with 100% of subjects with no prior knowledge.

The next section examines inter-subject agreement, and this will shed some light on the findings relating to the continuous variable scores.

### 5.1.7 Inter-subject Agreement

To get a better understanding of the reliability and validity of the statistical results, we evaluated how well the subjects agreed in their evaluation of the subtitles with respect to the four continuous variables. To calculate inter-subject agreement for comprehension, errors and style we used leave-one-out resampling.[96] For satisfaction scores we used the Kappa coefficient[97] to calculate inter-subject agreement. Kappa coefficient (*k*) is defined as:

$$k = \frac{p_a - p_r}{1 - p_r}$$

where $P_a$ is the proportion of times that the annotators agree, and $P_r$ is the probability that they would agree by chance. We follow Callison-Burch et al. (2007) when defining chance agreement $P_r$, and use 1/6 since there are six possible outcomes when rating the overall satisfaction with the subtitles. For $P_a$ we calculated the proportion of time they assigned identical scores to the subtitles.

The inter-subject agreement using the leave-one-out resampling technique for Corpus AM was **r = 0.32**, Corpus BM was **r = 0.69**, and Corpus CM was **r = 0.30**. These results can be interpreted the same as Kappa scores (see footnote 43) meaning Corpus AM score indicates *fair* agreement, Corpus BM score indicates *substantial* agreement, and Corpus CM score indicates *fair* agreement.

The inter-subject agreement using the Kappa coefficient technique for Corpus AM was **k = 0.39**, Corpus BM was **k = 0.35**, and Corpus CM was **k = 0.35**. These results indicate *fair* agreement for all three corpora when rating overall satisfaction.

[96] According to Lapata (2005:478) this technique is a special case of *n*-fold cross-validation (Weiss and Kulikowski 1991) and has been previously used for measuring how well humans agree on judging semantic similarity (Resnik and Diab 2000; Resnik 1999), adjective plausibility (Lapata and Lascarides 2003), and text coherence (Barzilay and Lapata 2005). For each of our corpora, the set of subjects' responses, including a score for comprehension, style and errors (*m*) was divided into two sets: a set of size $m - 1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the mean ratings of the former set with the ratings of the latter. This was repeated *m* times. Since we had 15 subjects, we performed 14 correlation analyses and report their mean.

[97] We use this technique for satisfaction scores as there are only 3 sets, with 15, 15 and 14 subjects respectively.

### 5.1.8 Summary of Statistical Tests

The results of the statistical tests indicate possible differences between the corpora in relation to the independent (number of SL repetitions, corpus size, corpus homogeneity) and dependent variables (intelligibility and acceptability):

- The perceived speed of the subtitles was deemed unsuitable by a higher percentage of subjects viewing Corpus BM subtitles, suggesting that the readability of Corpus BM is lower than Corpus AM or Corpus CM
- We saw earlier that 83% of Corpus BM subtitles match Corpus AM subtitles, which might indicate that a high number of subjects deemed the subtitles too fast due to other factors such as being unfamiliar with watching a movie using subtitles
- Corpus CM subjects noted the highest number of errors, and in particular a significantly different number of C1 errors than subjects from the other corpora. The increase in C1 errors could be related to the introduction of less genre-specific data, as 33% of Corpus CM subtitles do not exactly match subtitles in either Corpus AM or Corpus BM. However, the subjects rated the severity of C1 errors similar to Corpus BM subjects, meaning the observation of more C1 errors lowers the well-formedness of the subtitles, but not necessarily the intelligibility of the subtitles (cf. Flanagan 1994). From these results Corpus CM is deemed the least well-formed
- Corpus AM subjects noted a significantly lower number of C1, and a similar number of C2 and C3 errors as the other two groups of subjects. However, Corpus AM subjects rated the errors as the most annoying of all the subjects. This result combined with the previous one suggests Corpus BM subtitles are the most well-formed
- In addition Corpus AM subjects rated the style of the subtitles as the least appropriate in comparison with Corpus BM and Corpus CM. We have already mentioned how this result seems to contradict any preconceived ideas regarding the optimum performance of genre-specific corpora (high homogeneity)
- The combination of rating the errors as most annoying and the style as least appropriate suggests Corpus AM subtitles are the least acceptable of the three corpora

The statistical tests looking at additional independent variables of soundtrack language, linguistic background and prior knowledge suggest the following:

- If subjects have knowledge of the soundtrack language they perceive the subtitles as more comprehensible (comprehensibility scale) and more well-formed (error scale)
- Knowing the soundtrack language does not have an impact on how the subjects measure the appropriateness of style
- Subjects watching movie clips with an unknown soundtrack language rely more on the subtitles to understand
- Intra-corpus results for Corpus BM show that subjects are less likely to rely on the subtitles to understand when they have knowledge of the soundtrack
- Subjects with a linguistic background rated the subtitles as less comprehensible than subjects with no linguistic background
- There are significant inter-corpus differences when we consider either subjects with prior knowledge or subjects with no prior knowledge. Errors (no prior knowledge), style (prior knowledge) and overall satisfaction (prior knowledge) are all rated lower by Corpus AM subjects than the other two corpora. There are also significant intra-corpus differences for Corpus CM relating to overall satisfaction and Corpus AM relating to the categorical variables. Corpus CM subjects with prior knowledge consider the subtitles to be more satisfactory than subjects with no prior knowledge. Corpus AM subjects with prior knowledge consider the subtitles well-translated more often than subjects with no prior knowledge. This finding shows that prior knowledge has an impact on the well-formedness of the subtitles. However, judgements from Corpus AM subjects with no prior knowledge improve the readability results

We should also note that the inter-subject agreement in relation to the three continuous variables (comprehensibility, errors and style) is higher for Corpus BM than the other two corpora (*substantial*). Inter-subject agreement for overall satisfaction is the same for all three corpora (*fair*).

In the next section we look at qualitative data collected during the evaluation sessions, and examine whether the qualitative data can help us develop further observations made during the quantitative phase.

## 5.2 Qualitative Analysis

A common qualitative analysis technique adopted in this research is the coding of subjects' responses (Creswell 1998:140, Oppenheim 1992:266). Coding involves attaching labels to segments of data that describe what each segment is about (Charmaz 2006). The main aim of coding is to reduce the data, which "aids the organization, retrieval, and interpretation of data" (Coffey & Atkinson 1996:27). Following the coding process categories or sub-categories begin to emerge from the data. In the current study the primary source of qualitative data is the responses to open questions on the questionnaire and any additional comments that subjects included with closed questions or in conversation with the researcher. This study differs in two ways from studies where coding is normally used: firstly we do not have a 'large quantity' of data in comparison to say unstructured interview transcripts; and secondly, the main categories were already established prior to the coding, and relate to our four quality characteristics (comprehensibility, readability, style and well-formedness), plus an additional *overall* category described in Chapter 3. As previously outlined in Chapter 3, there are specific questions in the questionnaire within each of these categories. When conducting a qualitative analysis, we examine the subjects' responses within the five categories, assign codes to the responses, and then group these codes into sub-categories. Figure 5.4 overleaf presents an example of the coding structure.

**Figure 5.4:** Example of Readability (category) for Corpus AM. Categories are established *prior* to the interview, and sub-categories are established *following* the analysis

**Corpus AM: Readability (Category)**

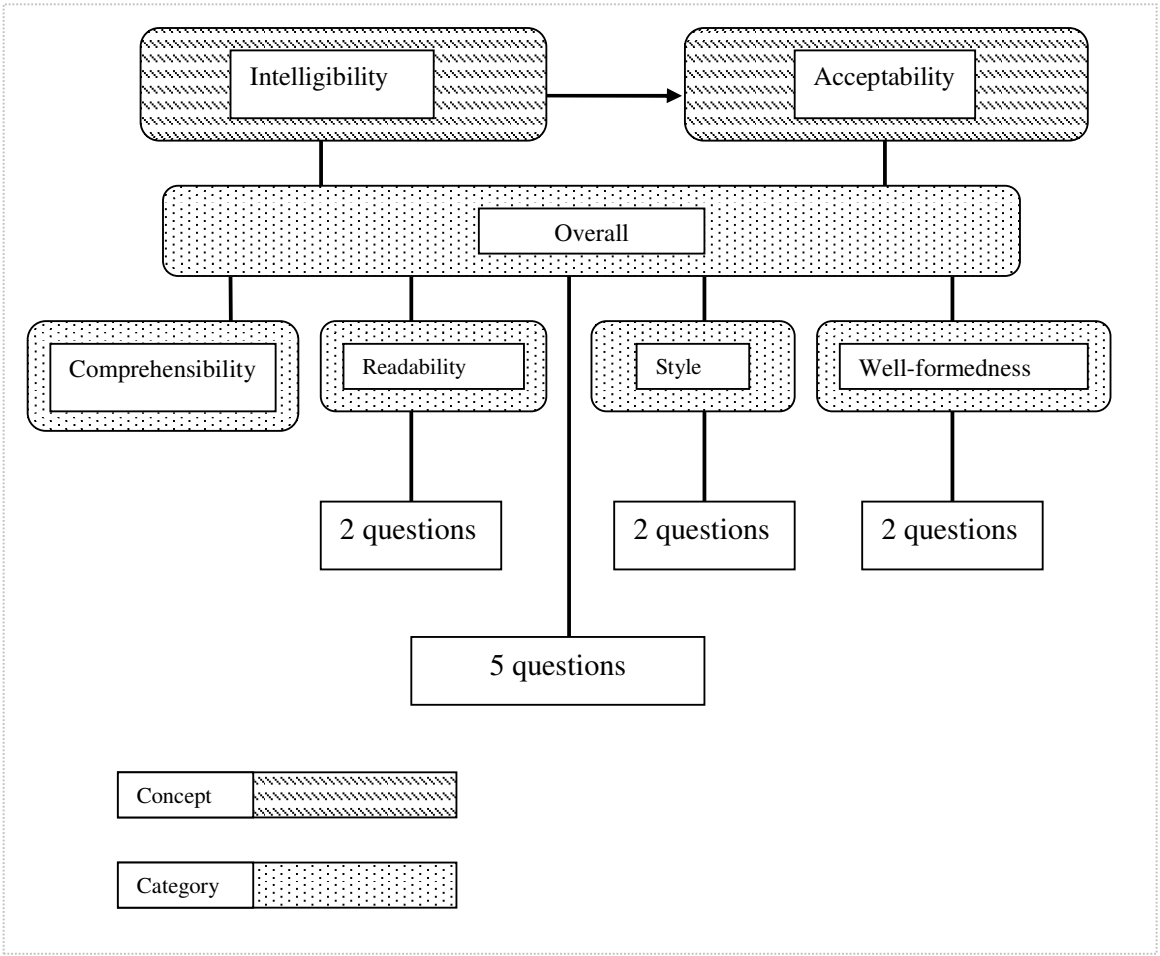
- **Comprehension issues (sub-category)**
  Movie clips contain incomprehensible subtitles (20)
  Speed of the subtitles was too fast to read them in full (4)
  Subtitles are distracting and do not help with comprehension (2)
- **Linguistic issues (sub-category)**
  Incorrect grammar (21)
  Incorrect word order (5)
  Unnatural direct or word-for-word translations (2)

We can describe the number of references within a single code as the *density* of the code. Although the density of a code is not necessarily an indication of its importance to quality characteristics we are measuring, dense codes are noteworthy given that they indicate opinions of the subjects which recur relatively frequently within the core categories (Dunne 2008). So from the example in Figure 5.4, the linguistic issues sub-category contains three codes, with a density of 21, 5 and 2 respectively. Perhaps a drawback of coding by the researcher is that it is a subjective and interpretive process, and another researcher could make different claims while using the same data set. However, subjectivity is an integral part of qualitative research and something which cannot be avoided in this study.

### 5.2.1 Categories and Sub-categories

Following an analysis of the qualitative data, sub-categories emerged from within the already established categories. We begin the discussion of our qualitative analysis by firstly referring to Figure 5.5. Here we outline the hierarchy of categories within the study and this is a useful guide to refer back to as we progress through the next sections.

**Figure 5.5:** Qualitative categories investigating the concepts of intelligibility and acceptability and the corresponding number of questions used to gather data




The *overall* category relates to questions that were gathered at the end of each evaluation session relating to the subtitles as a whole, and not looking at particular movie clips. The questions also did not relate specifically to any of the other four quality characteristics (categories), and information gathered from the *overall* questions could be used to identify the issues in any of categories. Even though the *overall* category is positioned higher up in the categories, we will deal with it last.

### Comprehensibility Category

We can see from Figure 5.5 that there are no qualitative questions within *comprehensibility*. The questions we asked the subjects were quantitative in nature, and therefore we did not gather any qualitative data specific to this category. We will observe, however, that some of the other categories incorporated qualitative data relating to comprehensibility.

### Readability Category

Within *readability*, subjects were asked one question which generated data to be grouped into sub-categories:

- During the movie clips did you notice any subtitles that seemed out of context?

If subtitles which are out of context appear on the screen, this has an impact on readability. The viewer is confused by the introduction of a subtitle that is not directly related to the image, and may have to re-read the subtitle, which is not always possible due to the time-restricted environment. From the quantitative data given in Table 5.25 below, all three corpora were identified as having produced subtitles out of context.

**Table 5.25:** Number and percentage of subjects who said that there were subtitles out of context

| | Corpus | | | |
|---|---|---|---|---|
| **Subtitle Out of Context** | **AM** | **BM** | **CM** | **Total** |
| **Yes** | 34 | 36 | 24 | 94 |
| % within Corpus | **37.8%** | **40.0%** | **28.6%** | |
| **No** | 56 | 54 | 60 | 170 |
| % within Corpus | **62.2%** | **60.0%** | **71.4%** | |
| **Total** | **90** | **90** | **84** | |

In addition, the qualitative data for this question tells us that subjects from all three corpora were able to recall some subtitles they thought were out of context: Corpus AM (17 subtitles), Corpus BM (12 subtitles) and Corpus CM (9 subtitles). We can see from these figures that the subjects recalled only 38 subtitles (tokens) out of a possible 94. Of the 38 they recalled, there are 26 different subtitles (types): eight of these were identified by at least two of the three groups, and three identified by all three groups (see Appendix J for a comprehensive list of the subtitles identified).

The data in Table 5.25 show Corpus BM was identified as having the highest percentage of subtitles out of context, followed closely by Corpus AM. The qualitative data reverse this result, with only 33% of Corpus BM subjects successfully recalling subtitles they deemed out of context, in contrast to 50% of Corpus AM subjects. Corpus CM has the lowest number of subtitles deemed out of context by the subjects. From these results we could say that the readability of Corpus CM subtitles is higher than the other two in terms of having the lowest number of subtitles deemed out of context, both in terms of quantitative and qualitative data.

Combining the quantitative and qualitative results for r*eadability*, we consider subtitle speed and subtitles deemed out of context: Corpus CM is considered to be the corpus that has the lowest number of subtitles out of context; it is also the corpus with the highest percentage of subjects who thought the speed of the subtitles was suitable. From these combined results we can claim that in terms of the readability of EBMT-generated subtitles, Corpus CM is ranked higher than the other two.

### Style Category

Within s*tyle*, subjects were asked two questions which generated data to be grouped into sub-categories:

- Did anything bother you about the subtitles?
- Did anything amuse you about the subtitles?

The two questions were intended to elicit responses about style, in the knowledge that subtitles particularly bothered or amused the subjects, given that subtitles may be highly readable but in an inappropriate style. Asking subjects if something bothered them elicited many negative opinions and asking them if something amused them elicited both positive and negative opinions in relation to style. When analysing the data collected for the style questions we ignored comments that specifically overlap with previous data collected from the other quality characteristics, for example grammar as something that bothered the subject, as this would be noted during a question on errors or overall dissatisfaction. We did this to try to avoid double-counting the data.

There were fifteen different points that bothered subjects across the three corpora: Corpus AM (12), Corpus BM (3) and Corpus CM (3) (see Appendix K for a

comprehensive list of the points identified). Four of the points mentioned relate to the use of an inappropriate style: first, the use of an adjective in a certain context. One example is when Harry says e*igenartig* when looking at the broom. The subject thought that he should have used *komisch* instead (Corpus AM).[98] Second, the translation of the surname *Susan Bones* as *Susan Knochen* (Corpus AM and Corpus BM. The original English was reproduced in subtitles from Corpus CM).[99] The *Susan Bones* 'error' could be flagged easily in the system as a proper name and therefore it would not be translated, keeping in line with all merchandise translations.

The third item concerns a conversation between Hermione and Harry:

**(Hermione)**

> **English**: You'll be ok Harry. You're a great wizard. You really are.
> **German EBMT**: Ihr verbüßt ok, Harry. Du bist ein großer Zauberer. Wirklich.

The German subtitle (from Corpus AM) uses the verb *verbüßt* and one subject thought this was very inappropriate and that the style was too high. This is an incorrect verb in this context, but we also mention it as the subject noticed the style element too.

One subject commented that it did not seem like a native speaker had translated the subtitles due to the incorrect style.

The remaining items of subtitles that bothered the subjects relate to the other three quality characteristics. The items relating to readability and well-formedness are accounted for elsewhere in the analysis. In relation to comprehensibility, subjects commented on the following issues that hampered comprehension:

---

[98] Interestingly enough *eigenartig* is the translation offered for *strange* on the purchased DVD, and it was also deemed suitable by other subjects.

[99] The translation of *Susan Bones* as *Susan Knochen* only bothered Corpus AM subjects. According to Brøndsted & Dollerup (2004:58), a Swedish translator reported that translators of the Harry Potter books had to sign a contract "agreeing to keep the original names, so Warner Brothers can distribute the films, computer games and other merchandise all around the world with the names everyone recognizes" (Fries-Gedin 2002 cited in Brøndsted & Dollerup ibid). The authors also point out that in the German translation of the *Harry Potter* books, there are only two name changes: firstly, Hermione Granger, a main character in the story, is naturalised in German making it easier to pronounce (Hermine Granger), but consequently making it harder to connect the name to Shakespeare, and secondly a 'descriptive name' Nearly-Headless Nick (Der Fast Kopflose Nick). All other names are transferred from the English originals.

- They could understand more from using the Dutch soundtrack rather than the subtitles (Corpus AM)
- There were some particular subtitles that could be considered 'strange' as the subjects could not understand them (Corpus AM)
- One subject only understood the meaning of the subtitle from using the English language soundtrack (Corpus AM)
- Often reading 'normal words' was problematic (Corpus CM)

All of these comprehensibility issues relate to characteristics of raw EBMT output, and these comments are not surprising.

When analysing the data on what amused the subjects about the subtitles, we once again ignored comments that specifically overlap with previous data collected from the other quality characteristics (see Appendix L for the complete table of items that amused the subjects). Fifteen different items amused the subjects, distributed as follows: Corpus AM (8), Corpus BM (5) and Corpus CM (2). Of these fifteen items, nine of them relate to the style of the subtitles. Five of the nine items were somewhat negative observations (Corpus AM (4), Corpus CM (1)), with the most negative items including unusual use of adjectives (*Wir wollen ein überzeugter, altmodischen, vermutlich rostig*), not being able to read any correct subtitles and the subtitles not capturing the mood of the conversation. A slightly less negative item includes '*I could imagine in many instances what the subtitle should have said!*' (Corpus BM). The remaining four items include positive comments regarding subtitle translation which incorporate references to style and other quality characteristics, including the appropriate use of kids' speak (readability, style and well-formedness – all three corpora), suitable subtitles describing the pictures by the stairs (comprehensibility, readability, style, well-formedness – Corpus BM) and once again a comment on the translation of *Susan Bones* as *Susan Knochen* (style – Corpus BM).

The quantitative results presented earlier showed that there was a significant difference between Corpus AM judgements and judgements from Corpus BM and Corpus CM in relation to something amusing in the subtitles (see Table 5.5). Of the five most negative comments, four of these refer to Corpus AM subtitles. There was also a significant difference between the style scores for Corpus AM and Corpus BM, with Corpus AM

subtitles rated as having the least appropriate style (see Table 5.6). Corpus BM subtitles were rated as exhibiting the most appropriate style, but there were not many examples given by subjects in support of this when asked to expand on these ratings. From the combined results we can say that Corpus AM subtitles have the least appropriate style, but we are unable to make any claims regarding differences between Corpus BM and Corpus CM for style.

### Well-Formedness Category

Within *well-formedness*, subjects were asked two questions which generated data to be grouped into sub-categories:

- Are the EBMT subtitles well-translated[100] for use on a DVD if the viewer does not understand the soundtrack language?
- Can you recall any well-translated subtitles?

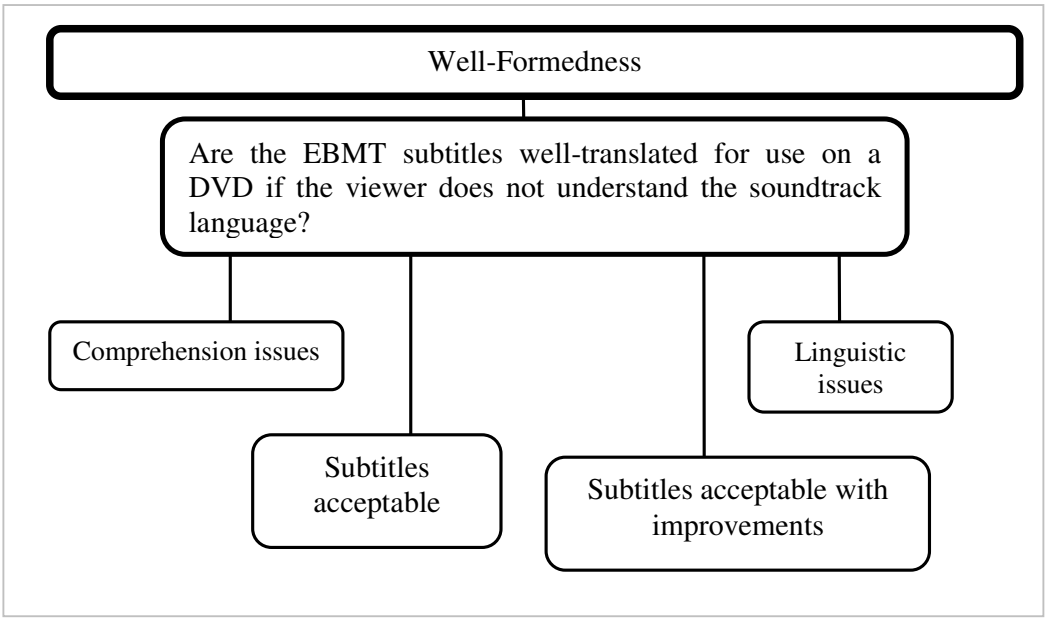
Within each corpus the data were coded[101] and these codes were grouped into sub-categories. The sub-categories are presented in the next few sections, and the results show that some sub-categories and codes were common to all three corpora.

[100] Well-translated in this thesis is taken to mean well-formed, and follows the definition of well-formedness as outlined in the Introduction. We believed that subjects would understand well-translated more than well-formed, especially if they had no formal language training. As the subjects only had access to the translated subtitle, they focussed their response on this translation, and did not refer to the original subtitle.

[101] Given the large number of comments from subjects, the codes were formulated by the researcher and it was noted how many references to each code were in the data.

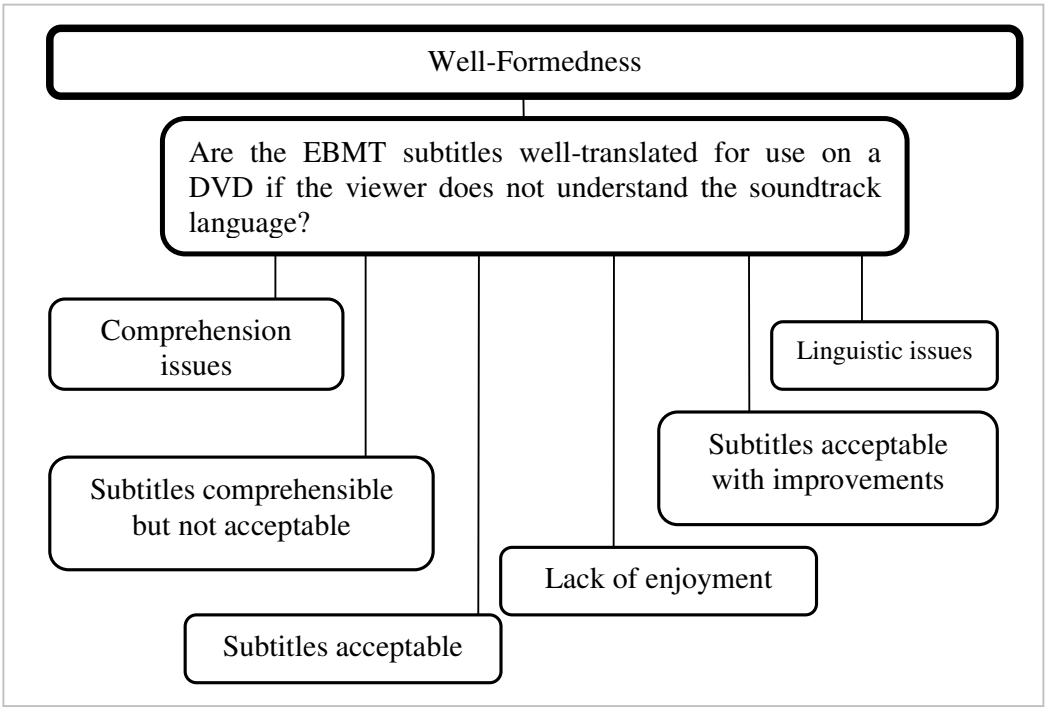
### Corpus AM

After analysing the data, four sub-categories emerged (see Appendix M for a list of codes for all three corpora):



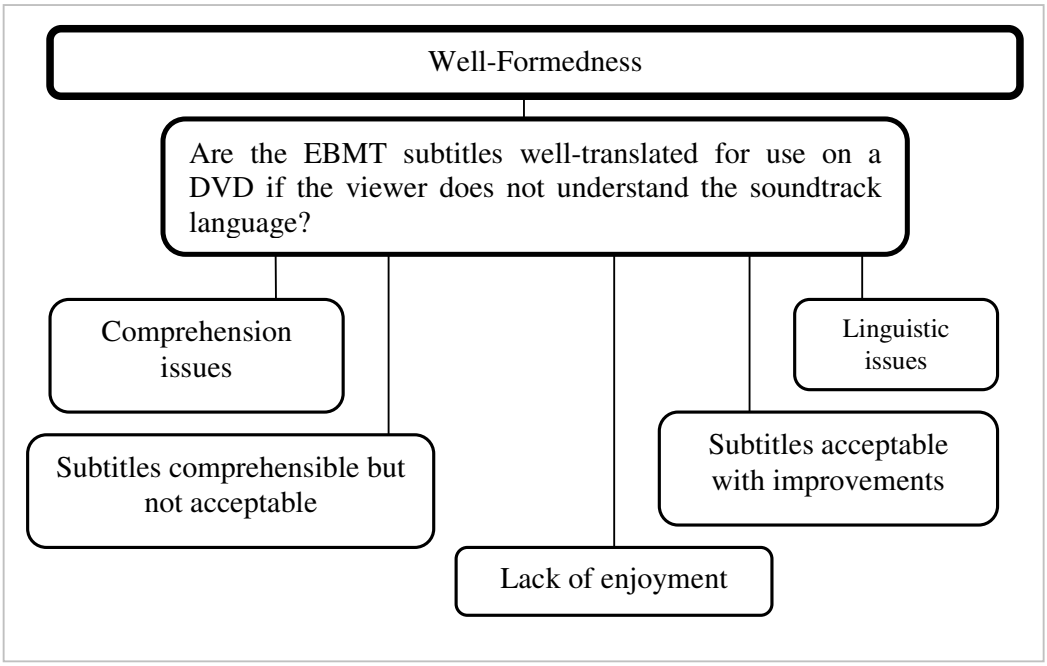

### Corpus BM

After analysing the data, six sub-categories emerged for Corpus BM, four of which are the same as Corpus AM:



### Corpus CM

After analysing the data five sub-categories emerged, which are identical to five of the six that emerged from Corpus BM and three of the four that emerged from Corpus AM:




From the results we identify six sub-categories between the three corpora: comprehension issues, linguistic issues, subtitles acceptable with improvements, subtitles comprehensible but not acceptable, lack of enjoyment and subtitles acceptable. Table 5.26 presents the number of positive and negative codes per corpus within these sub-categories that are relevant to the well-formedness of the subtitles.

**Table 5.26:** Inter-corpus results for well-formedness references (both positive and negative codes)

| | | **Corpus AM** | **Corpus BM** | **Corpus CM** |
|---|---|---|---|---|
| **Positive** | Number of positive codes within the sub-categories *Comprehension issues*, *subtitles acceptable with improvements* and *subtitles acceptable*, all of which are relevant to well-formedness | 9 (17%) | 16 (26%) | 10 (14%) |
| **Negative** | Number of references relating to codes within the sub-categories *Linguistic issues, comprehension issues,* and *lack of enjoyment*, all of which are relevant to well-formedness | 44 (83%) | 45 (74%) | 61 (86%) |

We can see from these results that Corpus CM subtitles received the highest number of negative responses, which suggests these subtitles are the least well-formed.

The second open question within the w*ell-formedness* category asked the subjects if they could recall any well-translated subtitles in the clips they viewed. Table 5.27 presents the results.

**Table 5.27:** Number of times subjects could recall well-translated subtitles across the six movie clips

| | Corpus | | | |
|---|---|---|---|---|
| **Well-Translated Subtitles** | **AM** | **BM** | **CM** | **Total** |
| **Yes** | 61 | 54 | 49 | 164 |
| % within Corpus | **67.8%** | **60.0%** | **58.3%** | |
| **No** | 29 | 36 | 35 | 100 |
| % within Corpus | 32.2% | 40.0% | 41.7% | |
| **Total** | **90** | **90** | **84** | |

Referring to the quantitative data in Table 5.27, Corpus AM subjects accounted for the highest percentage of well-translated subtitles observed, followed by Corpus BM and Corpus CM respectively. Subjects were then asked to specify the well-translated subtitles. Of the numbers of subtitles given in Table 5.27 above, Corpus CM subjects were able to recall the highest percentage of examples (43%), followed closely by Corpus BM subjects (39%) and then Corpus AM subjects (30%) (see Appendix N for the complete list of well-translated subtitles recalled), the reverse order of the data presented in Table 5.27.

Reflecting on the quantitative data for well-formedness, we noted the following: there was no significant inter-corpus difference between the number of subjects who thought the subtitles were well-translated and the number who thought the subtitles were not well-translated for use on DVD. There was also no significant inter-corpus difference between the number of subjects who noted well-translated subtitles and those who did not. In addition there was no significant inter-corpus difference in mean error scores: 2.6 (AM), 2.8 (BM), 2.8 (CM). Subjects viewing subtitles from all three corpora noted errors in the subtitles (98.9% (AM), 100% (BM), 97.6% (CM)), but subjects viewing subtitles from Corpus CM noted the highest percentage of C1 errors (least effect on intelligibility) and the lowest percentage of C3 errors (most effect on intelligibility).

The qualitative data (Table 5.26) showed Corpus CM subtitles were not considered as well-translated as subtitles from either Corpus AM or Corpus BM for use on DVD.

In summary, Corpus CM subjects noted the lowest number of 'serious' errors, but a significantly higher number of 'least serious' errors. They noted the lowest number of well-translated subtitles, but of these subtitles, they were able to recall the highest number of examples. The fact that Corpus CM subjects noted such a significant number of C1 errors and fewer well-translated subtitles suggests Corpus CM subtitles are less well-formed than subtitles generated by either of the other two corpora.

Looking at the results for the other two corpora, we can suggest that Corpus BM subtitles are more well-formed: error scores for Corpus BM are higher than Corpus AM (meaning Corpus BM subtitles are less annoying than Corpus AM subtitles); Corpus BM received a higher number of positive comments relating to well-translated subtitles (and only one more negative comment); and Corpus BM subjects could recall a higher number of well-translated subtitles than Corpus AM subjects.

### Overall Category

Within the *overall* category, subjects were asked four questions after viewing all six clips:

- Did you notice any repeated subtitles throughout the six clips?
- Are the subtitles more acceptable on clips with a Dutch language soundtrack or with an English language soundtrack?
- After viewing all six clips, on a scale of 1-6 (1 being very dissatisfied and 6 being very satisfied) how would you rate the subtitles?
- Do you have any overall comments regarding your satisfaction/dissatisfaction with the subtitles?

We also introduce an additional question in this category. During the evaluation sessions subjects were asked one question after each of the three clips with the unknown language soundtrack (Dutch):

- Would you use these subtitles on a DVD if you did not understand the soundtrack language?

This individual question is considered in the *overall* category as it is not specific to any of the four quality characteristics, but rather to the overall acceptability of the subtitles when the subject does not understand the soundtrack.[102]

We look firstly at the results of whether subjects noticed any repeated subtitles (Table 5.28).

**Table 5.28:** Percentage of subjects (per corpus) who noticed any repeated subtitles (same German translation) throughout the movie clips

| | **Yes** | **No** |
|---|---|---|
| **Corpus AM** | 13% | 87% |
| **Corpus BM** | 13% | 87% |
| **Corpus CM** | 21% | 79% |
| **What were the subtitles?** | **Nicht bummeln** – keep up<br>**Was ist los**? – What's wrong/what is it?<br>**Komm schon** – come on (one person)<br>**Kommt schon** – come on (group)<br>**eigenartig** – strange/curious<br>**Hilfe** – help<br>**Troll im Kerker** – troll in the dungeon | |

There were only a few repeated subtitles included in the six clips, and 16% of all subjects noticed these repetitions. Some of the repeated subtitles could be translated in slightly different ways, using translations from the corpora (Table 5.29):

**Table 5.29:** Alternative translations within the corpora for the repeated subtitles

| | | Alternative translations | | |
|---|---|---|---|---|
| nicht bummeln | → | nicht trödeln | | |
| was ist los? | → | was ist denn? | → | was ist? |
| Komm schon | → | na komm | → | na los |
| eigenartig | → | merkwürdig | | |

[102] There are similarities between this question and the one within the well-formedness category. However, this question is not specific only to the well-formedness of the subtitles. It aims to establish whether the subject would use subtitles of a similar standard to view an entire movie, and removes any influence from the known source language (English) when considering the overall acceptability of the subtitles.

However, the EBMT system chose the same translation on each occasion. During a discussion at the Languages and the Media 2006 conference, subtitlers from Scandinavian countries were concerned that introducing MT would mean introducing repetitiveness and eliminating literary creativity often associated with creating subtitles (see Volk 2008:208). Given that the number of repeated subtitles in this study represents only a small percentage (6%) of the overall number presented in the movie clips it is not possible to make a general claim about how viewers might react to repeated translations throughout an entire movie. But this result offers some weight to the argument that if MT was used for producing subtitles, therefore introducing an element of repetition, the viewers would not necessarily find this a distraction or feel they were being offered a sub-standard service. Later in section 5.2.2 we make further reference to some of these repeated subtitles, when we discuss subtitles that polarised the subjects.

We saw earlier in section 5.1.6 that the statistical tests returned a significant result for soundtrack language having an impact on two of the three continuous variables, namely comprehensibility and errors, within the corpora. These findings suggest that subjects consider the subtitles more comprehensible and more well-formed when they understand the soundtrack language. However, we can see from the results in Table 5.30 below that over half of the subjects (52%) across the three corpora considered the subtitles to be of the same standard overall, irrespective of the soundtrack language. That said, we should comment that of the subjects who noticed a difference, 69% thought the subtitles were better on the clips with the English soundtrack. This result is at variance with findings from Armstrong et al.'s (2006c) pilot study, in which subjects were more accepting of subtitles when they viewed movie clips with an unknown language soundtrack (Japanese) than with a known language soundtrack (English).

**Table 5.30**: Results for whether subjects noticed any difference in the quality of the subtitles depending on the language of the soundtrack

| | **Higher quality subtitles on English language soundtrack clips** | **Higher quality subtitles on Dutch language soundtrack clips** | **No difference in quality** |
|---|---|---|---|
| **Corpus AM** | | | |
| Starting Dutch[103] | 1 | 0 | 6 |
| Starting English | 1 | 0 | 7 |
| **Total AM** | **2 (13.3%)** | **0** | **13 (86.7%)** |
| | | | |
| **Corpus BM** | | | |
| Starting Dutch | 1 | 2 | 4 |
| Starting English | 3 | 0 | 5 |
| **Total BM** | **4 (26.6%)** | **2 (13.4%)** | **9 (60%)** |
| | | | |
| **Corpus CM** | | | |
| Starting Dutch | 0 | 3 | 4 |
| Starting English | 5 | 0 | 2 |
| **Total CM** | **5 (35.7%)** | **3 (21.4%)** | **6 (42.9%)** |

We then asked subjects to rank the subtitles on a scale of 1-6, (1 being very dissatisfied and 6 being very satisfied). We ran a statistical test on the scores, and there was no significant difference between the mean scores for the corpora. We can see from Table 5.31 that Corpus CM has the highest mean score (2.8).

**Table 5.31:** Mean scores for overall satisfaction with the EBMT subtitles

| **Corpus** | **Mean** | **N** | **Std. Deviation** |
|---|---|---|---|
| AM | 2.3000 | 15 | .84092 |
| BM | 2.6000 | 15 | .91026 |
| CM | **2.8571** | 14 | .86444 |

Earlier when we tested whether prior knowledge (PK) had an impact on the continuous variables there was a strong effect for PK on the overall satisfaction scores, both inter-corpus (Corpus AM and Corpus CM) and intra-corpus (Corpus CM). As mentioned previously, the data showed that Corpus CM subjects with PK rated overall satisfaction

[103] Starting Dutch and Starting English refers to the soundtrack language of the first clip in the sequence of six. The two languages alternated for the six clips.

significantly higher than Corpus AM subjects with PK. For the intra-Corpus CM result we must consider PK as a contributing factor to the satisfaction rating.

The final question following the six clips relates to the scale question, asking subjects if they had any overall comments regarding their satisfaction/dissatisfaction with the subtitles (see Appendix O for a full list of comments). From the data eight sub-categories emerged, and we used the same sub-category groupings for each of the corpora:

- Helpful subtitles
- Unsuitable for learners of German
- Machine Translation and Post-editing
- Bad quality subtitles
- Using prior knowledge to understand the movie
- Unsatisfactory subtitles
- Unsuitable subtitles for a commercial DVD
- Dubbing versus subtitling

For all three corpora the sub-category *Helpful subtitles* is considered positive. The sub-category *Machine Translation and Post-editing* is considered a positive sub-category in Corpus CM and a negative sub-category in Corpus AM, and there are no references to it in Corpus BM. All of the other six sub-categories are considered negative. The sub-category dubbing versus subtitling refers to a subject's comment about always receiving the correct information when a movie is dubbed instead of subtitled. This idea contrasts with what has been reported in the literature (cf. section 1.2.1), which shows that viewers of dubbed material are more vulnerable to manipulation and censorship than subtitling viewers, and subtitlers are more vulnerable to criticism from viewers (cf. Díaz Cintas & Remael 2007). This view, however, shares similarities with Luyken et al.'s contention that dubbing is a 'faithful translation' of the original track (cf. section 1.2.1), although from the viewers' perspective, dubbed versions are presented as faithful translations. We are unable to use the findings from this study to argue in favour of either side, and it is a topic of future research not directly related to the current work.

For each of the corpora (see Figures 5.6-5.8), the positive categories are given on the left-hand side, while the negative categories are on the right-hand side, and the number of comments per category is also presented in parentheses. We omitted any sub-categories which contained no references.

**Figure 5.6:** Comments relating to overall satisfaction with the quality of Corpus AM subtitles

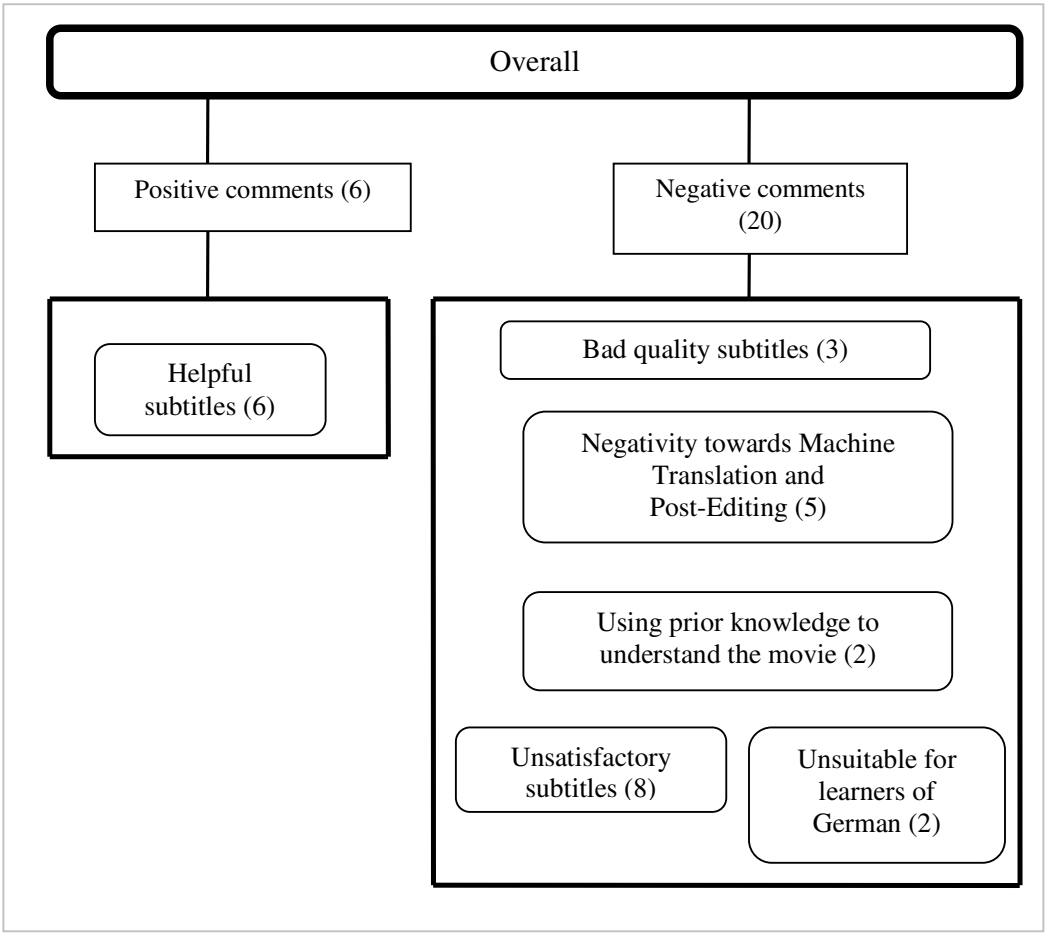


**Figure 5.7:** Comments relating to overall satisfaction with the quality of Corpus BM subtitles

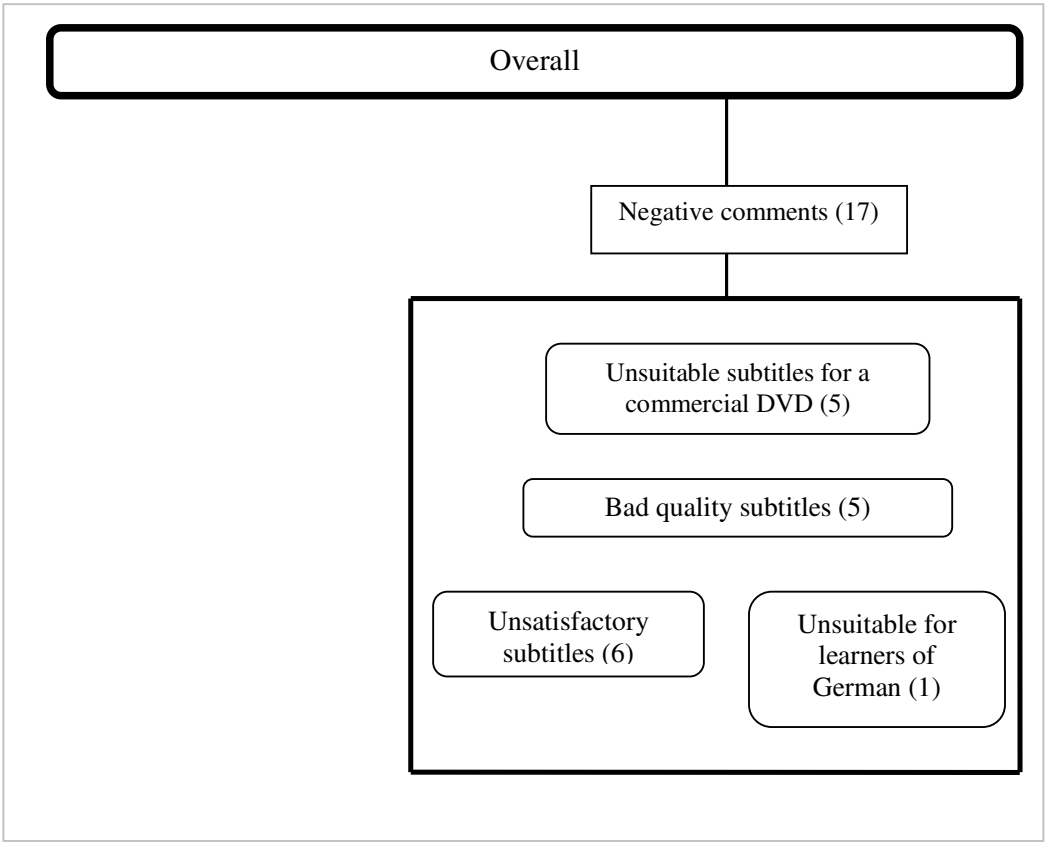
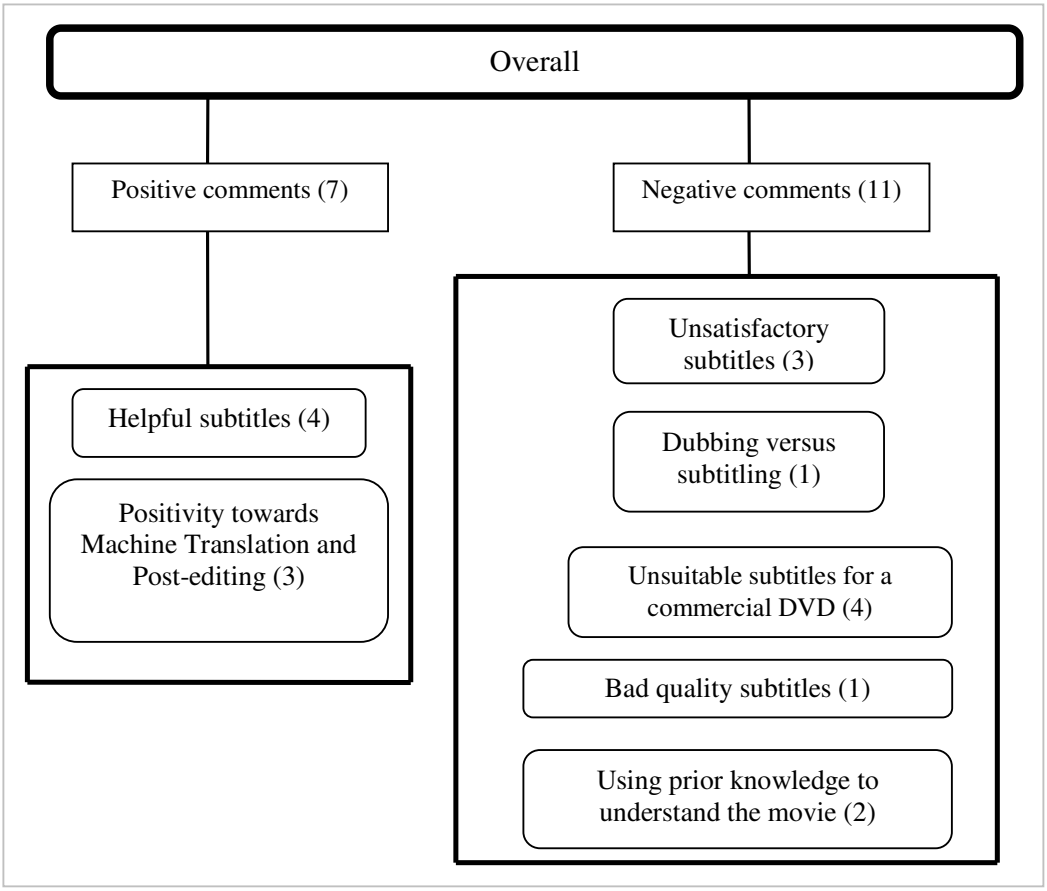


**Figure 5.8:** Comments relating to overall satisfaction with the quality of Corpus CM subtitles



Looking at Figures 5.6-5.8 we can see that for all three corpora, there were more negative than positive responses and Corpus BM had no positive responses. We note that Corpus CM received the highest number of positive responses and the lowest number of negative responses. However, the number of responses is not enough evidence to support Corpus CM as having generated the most satisfactory subtitles overall. This question is also difficult to analyse, given that subjects were asked for additional comments on satisfaction. This can result in subjects giving a few examples, on the one hand, or subjects giving only one answer, on the other. This means that we cannot compare numerical references to support any claims, unlike questions that generate a response from each subject.

The responses presented here within the *overall* category refer to problems associated with all four quality characteristics: comprehensibility, readability, style and well-formedness. The results have shown that the subtitles need to be improved in each of these areas if they are to be accepted as subtitles on a commercial DVD, a point which is further supported by the following subjects' opinions[104] on the matter:

> Basically yeah I was happy overall…happy with the understanding, but if I went to the cinema and had these [subtitles] I would be kinda pretty disappointed **(Subject C, Corpus AM).**

> Like I said short sentences they were translated pretty well, like yeah um, yeah they were right sentences, but kinda 'cause it just wasn't that complicated, but longer sentences sometimes, when the words could have different meanings, eh it was the wrong meaning **(Subject AB, Corpus AM).**

> I, yeah I would say it was 4 [overall satisfaction rating] because it is still possible to understand the movie and still yeah, you could still watch the movie and enjoy it, but since it's a film with lots of details, yeah, it wouldn't be really such a fine thing to watch it this way, some better translations come straight to my mind, so that kinda makes it weird as well, because you think why didn't they put it that way **(Subject AJ, Corpus BM).**

[104] The reader is reminded that these subjects are non-native speakers of English.

During the evaluation sessions, after the subjects viewed each of the three clips with the unknown language soundtrack they were asked whether they would use the subtitles shown for an entire movie if they did not understand the soundtrack. We analysed these qualitative data by putting the responses into different sub-categories within the groupings of *yes*, *maybe* and *no*. Firstly, Table 5.32 gives a brief quantitative overview of the subjects' responses.

**Table 5.32:** Responses to question whether subjects would use similar subtitles on a DVD with an unknown language soundtrack

| **Would you watch an entire movie with the German subtitles provided?** | **Yes** | **No** | **Maybe** | **Total** |
|---|---|---|---|---|
| **Corpus AM** | 38% (17) | 51% (23) | 11% (5) | **45** |
| | | | | (15 subjects x 3 clips) |
| **Corpus BM** | 20% (9) | 69% (31) | 11% (5) | **45** |
| | | | | (15 subjects x 3 clips) |
| **Corpus CM** | 40% (17) | 55% (23) | 5% (2) | **42** |
| | | | | (14 subjects x 3 clips) |

We can see from the quantitative results that for all corpora, subjects said they would not use the subtitles generated by the EBMT system in over half of the cases presented. The data also show that subjects from Corpus AM (38%) and Corpus CM (40%) are more likely to use the subtitles (based on findings from the three clips with the unknown language soundtrack), compared with the subjects from Corpus BM (20%). This contrasts with the earlier finding during the prospective phase where Corpus BM received the highest number of 'acceptable' responses for the EBMT-generated translations when the evaluation was conducted in a non-AVT environment. When analysing the data collected within the w*ell-formedness* category, we noted that subjects from all three corpora believed the subtitles are unacceptable from a commercial viewpoint, but that the subtitles are still acceptable in certain contexts (see section 5.2.1).

Figures 5.9-5.11 overleaf present qualitative data for the main reasons subjects might or might not use the EBMT-generated subtitles (yes groupings on the left-hand side,

maybe groupings in the middle and no groupings on the right-hand side). Subjects' judgements are connected to all four quality characteristics, with lack of well-formedness being the most dominant (see Appendix P for a full list of reasons per corpus and sub-category). The Figures show that a number of the sub-categories overlap between the corpora, and as we progress from Corpus AM to Corpus CM the number of sub-categories increases in the *maybe* and *no* groupings. However, the number of codes does not necessarily increase in the same way.

**Figure 5.9**: Corpus AM results for whether the subjects would use the subtitles if they did not understand the soundtrack

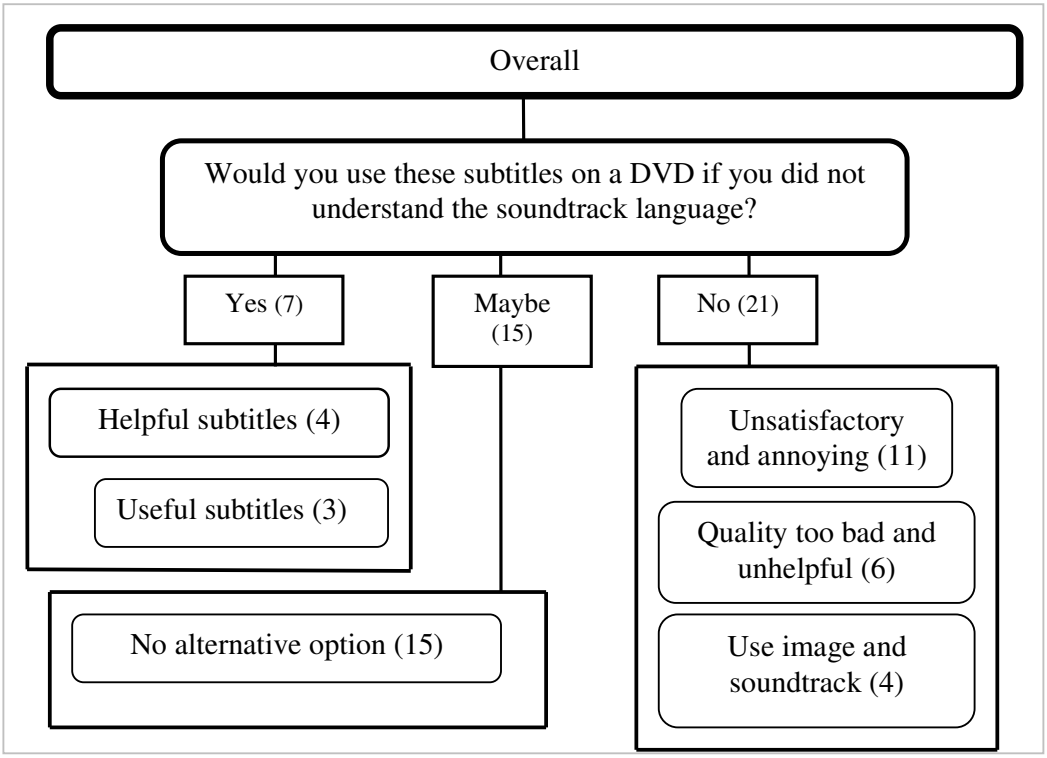


**Figure 5.10**: Corpus BM results for whether the subjects would use the subtitles if they did not understand the soundtrack

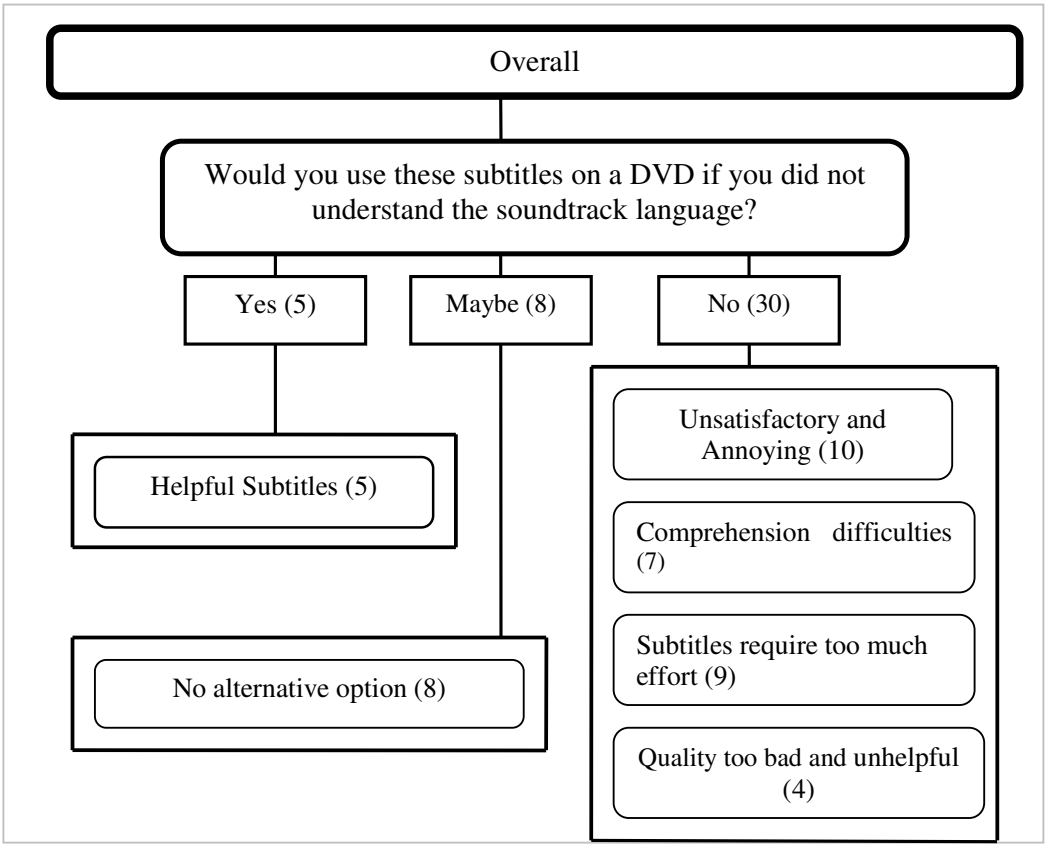
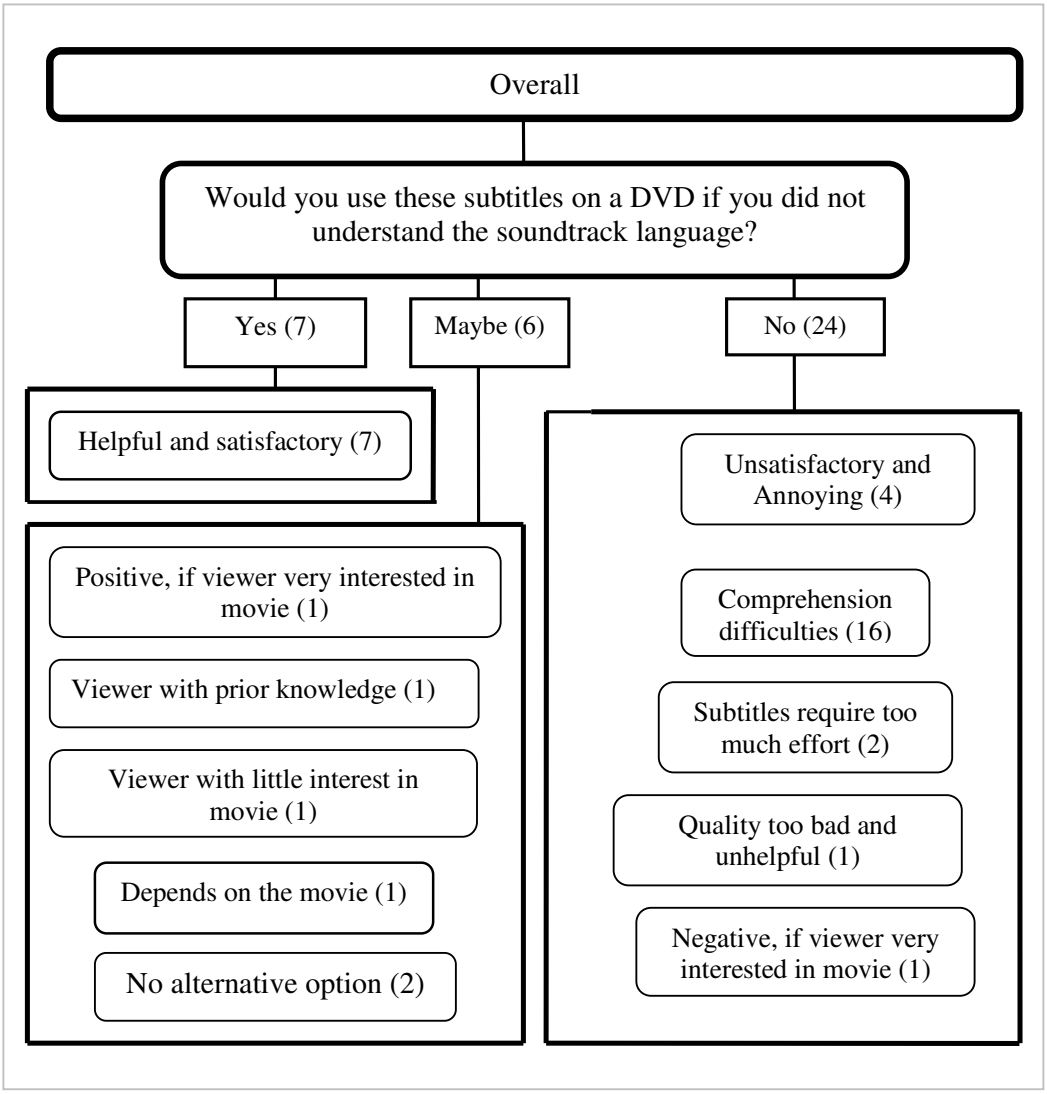


**Figure 5.11:** Corpus CM results for whether the subjects would use the subtitles if they did not understand the soundtrack



These qualitative data are further supported by subjects' comments on using the subtitles to watch a movie, with a mix of positive, negative and not sure responses. Subjects D and T were happy with some of the subtitles, but would not consider using this quality of subtitles on a commercial DVD:

> Yeah, you could use these subtitles to watch a movie, but there are still some errors, so yeah they would have to be improved before putting them on DVD or something like that **(Subject D, Corpus AM).**

> Some were ok, and then others were really, really, really bad. None of them were good enough to really be on a DVD that you want to sell…and because you can't really enjoy watching the movie, the thing is it would be like something to inform you, like that would be ok, but not to enjoy **(Subject T, Corpus CM).**

Subject H shows her dissatisfaction with the subtitles, but she would use the subtitles if she wanted to watch the movie and did not understand the soundtrack. Likewise for Subject E, even though he found some of the subtitles really confusing, he would rather use the subtitles than not use them:

> If there was no other option, then yeah I would use them, because I could get the gist from the subtitles. But really I would prefer not to watch them like, they could get annoying I'd say **(Subject H, Corpus AM).**

> I found some of the subtitles really helpful, like I would watch a movie with them, but then some of the others are really confusing, and I would have to turn off the movie. Some of the subtitles are borderline, but yeah I think I would rather use them than not use them. Perhaps it's easier to say yes [that the subtitles are acceptable] if you don't get the language like in this clip, if you don't really speak Dutch, then you don't know the language, and you are more concentrating on actually getting the content, it's better this way, but it could, yeah, the subtitles could still be improved **(Subject E, Corpus AM).**

Lastly, Subjects AQ and AH think they would use the subtitles provided on a DVD, and subject AF is almost sure he would use them, but notes they need improving:

> Well, I think I would because a combination of image and subtitle was good enough to get a meaning of the film and if you don't understand one or two subtitles, who cares? **(Subject AQ, Corpus CM)**
>
> Yeah it would be understandable, short sentences mostly yeah, but this is difficult to say if they are acceptable for a purchased DVD, yeah maybe **(Subject AH, Corpus BM)**
>
> 80% yes, em, yeah they could be used on a DVD, but probably need improvements, em **(Subject AF, Corpus BM).**

We will summarise the findings from the *overall* category. In relation to the question about whether subjects noted repeated subtitles, results showed that of the repeated subtitles presented to the subjects, only a small number actually noticed this. This implies that the repetition did not have a negative effect on the subjects, but as the data set is very small, this would have to be tested on a larger scale.

There was no significant difference between the satisfaction scores for the three corpora (but Corpus CM exhibited the highest mean score for satisfaction). The qualitative data for satisfaction (Figures 5.6-5.8) showed Corpus AM had the highest number of negative comments (20), followed by Corpus BM (17) and then Corpus CM (11). Corpus CM also received the highest number of positive comments of the three corpora. However, as we noted in section 5.1.7, the inter-subject agreement scores for satisfaction for all corpora were only deemed *fair*. The results could suggest that Corpus CM subtitles are more satisfactory overall than the other two corpora, but given the lack of significant quantitative findings and the fact that all three corpora received a *fair* inter-subject agreement rating casts some doubt on the validity of these results. Therefore the findings for overall satisfaction are not taken into consideration when we measure the intelligibility and acceptability of the subtitles.

When asked if the subjects would use similar subtitles to watch an entire DVD, Corpus BM fared the worst (30 no responses), followed by Corpus CM (24 no responses) and then Corpus AM (21 no responses). However all three corpora also received both *yes* and *maybe* responses. These findings highlight that the subtitles generated by the EBMT system can serve only as 'draft' versions of the final product, and would not be accepted on a commercial DVD in this form. However, this question also generated

positive responses to the subtitles and confirmed findings by Armstrong et al. (2006c) regarding the acceptability of draft subtitles in certain contexts, including online content and free downloads of subtitles.

In the next section we discuss additional themes to the ones we were specifically examining that emerged during the data analysis phase.

### 5.2.2 Additional Themes

During data analysis three themes emerged which need further comment.

#### Subtitles that polarise subjects

Some of the questions in the evaluation sessions asked subjects to recall specific subtitles in relation to well-formedness and style, including subtitles that seemed out of context, particularly well-translated subtitles and subtitles that bothered or amused subjects. We present two subtitles that polarised subjects, who either strongly agreed or strongly disagreed with the translation.

In the 'particularly well-translated' section, the subtitle (DE) *Voll krass!*[105] = (EN) *Wicked!* was mentioned nineteen times between the three corpora as 'a very appropriate thing for a young boy to say' and 'very colloquial'. In contrast to this, four subjects said this subtitle was 'out of context', 'inappropriate for a young boy to use' and 'not a suitable translation for the English subtitle *wicked*'. A suggested translation was '*gut'*, or *'toll'*. A second subtitle, (DE) *nicht bummeln* = (EN) *keep up* was recognised as one which was 'very appropriate, very correct German' within the context by ten subjects, and yet identified by two subjects as 'very inappropriate' in the context it was used. A third subject also suggested *bummeln nicht* would be more appropriate and natural sounding, but that they would accept the EBMT translation nevertheless. These examples show that with every evaluation involving human input there will never be 100% agreement.

[105] This was the subtitle translation provided on the purchased DVD.

### Subtitles that are not recognised, and therefore perceived as incorrect

Two more subtitles that caused disagreement were ones that contained *Harry Potter* specific terminology: (DE) *wutschen und wedeln* = (EN) swish and flick and (DE) *Eulerei* = (EN) owlery. 'Swish and flick' is the term used to describe the motion of using a magic wand to cast a spell. The second term 'owlery' describes a tower at the school where the post-owls live. They fly in and out of the tower delivering messages for the staff and students. There is no such word in the English language and it was created purely for the *Harry Potter* series. The German translation for this term is *Eulerei.* The EBMT-translated subtitles that include these terms use the same standard German translations offered in the books and in the subtitles on the official DVD. Six of the subjects who commented on these 'errors' had already read some of the books and had seen some of the movies. Thus we observed that subjects who did not recognise a term and therefore did not fully understand it, assumed that it was an error in the generation of the subtitles. This resulted in some subjects giving low scores to style and errors. They did not recall these unknown terms from their PK and therefore did not skew the data in a positive way, given that they were 'exposed' to the test material in advance of the study.

### Subtitles that elicit low scores, but are nonetheless acceptable

The final theme which emerged from the qualitative analysis concerns apparently conflicting opinions within a single evaluation session. For some clips subjects' ratings for comprehension, errors and style, and overall satisfaction are quite low (ranging between 1 and 3). However, when asked if they would use subtitles of a similar standard to watch an entire movie (after each of the clips with the unknown language soundtrack), their response was positive. In the next few paragraphs, we present some examples.

Subject AQ rated two clips 3 for comprehension, 2 for annoying errors and 3 for overall satisfaction. When asked if he would use subtitles of a similar standard to watch a movie on DVD he said:

> **(Clip 1)**
> Yeah I would watch the film, I could, yeah
> I could understand it, so I would use them **(Subject AQ)**.
>
> **(Clip 2)**
> Yeah I would, because you get most of the meaning,
> so they're ok **(Subject AQ)**.

Subject C rated one clip 2 for comprehension, 2 for annoying errors and 3 for satisfaction. When asked if she would use the subtitles of a similar standard to watch a movie on DVD she replied:

> If I didn't have a choice I would use them to watch a movie, I could understand, I was basically happy with the subtitles overall, for understanding purposes, but I would be disappointed if these were offered in the cinema **(Subject C)**.

Finally Subject AG gave one clip 2 for comprehension, 1 for annoying errors and 1 for satisfaction. She would, however, use the subtitles to watch a movie if these subtitles were offered. She gave a second clip 3 for comprehension, 3 for annoying errors and 3 for satisfaction, and added:

> I would use them if I had to, like, they could be used to understand some of the movie **(Subject AG)**.

She rated a third clip 3 for comprehension, 3 for annoying errors, 2 for style (register too high) and 3 for satisfaction. However, once again, she said she would use the subtitles to watch a movie, as she could understand what was meant by them.

This theme raises the fact that acceptability is a very difficult characteristic to measure. On the one hand subtitle users are very critical of features such as incorrect grammar, word order and style, yet on the other hand, they are accepting of these subtitles in some contexts. It is almost as if they say one thing and do another. Another factor we have to consider is that German-speaking countries are usually termed 'dubbing countries' and the viewers in this study would not have been exposed to subtitles to the same extent as, for example, viewers from Scandinavian countries. This can have a negative effect on the results for two reasons: firstly, if a person is simply not used to watching subtitles, they might find it more difficult to evaluate them; secondly, if a viewer has a negative

attitude towards subtitling in advance of the evaluation (which is the case for some subjects in this study), this negativity can come through in their judgements and could possibly bias the results.

As a result of this observation we wanted to examine whether subjects, who said at the beginning of the evaluation session that they did not like watching subtitles, rated the continuous variables (comprehensibility, errors and style) significantly lower than subjects who did not explicitly convey any negative attitudes towards subtitles. Table 5.33 below presents the mean scores from (1) subjects who said they do not like subtitles, and (2) the mean scores for all subjects. After conducting an independent t-test to compare two scores, none of the mean scores were significantly different for any of the independent variables. This means that subjects' negative attitudes did not have an impact on any of the findings presented.

**Table 5.33:** Mean scores of all subjects compared to the mean scores of subjects who do not like subtitles

| | | Mean Scores | | |
|---|---|---|---|---|
| | | **Comprehensibility** | **Style** | **Errors** |
| **Corpus AM** | 1 | 3.00 | 3.16 | 2.83 |
| | 2 | 3.10 | 3.35 | 2.63 |
| **Corpus BM** | 1 | 3.08 | 3.83 | 2.91 |
| | 2 | 3.35 | 3.87 | 2.80 |
| **Corpus CM** | 1 | 3.50 | 3.83 | 2.55 |
| | 2 | 3.17 | 3.61 | 2.82 |

1 = Negative towards subtitles 2 = All subjects

### 5.2.3 Automatic Metrics and Subtitling

MT evaluation is usually conducted using automatic metrics with BLEU being one of the most commonly used in MT research communities, having been described as the "*de facto* standard in machine translation evaluation" (Callison-Burch et al. 2007). The current study is predominantly a human evaluation study (retrospective phase), coupled with a corpus-analysis study to investigate factors that could influence the corpus used to train an EBMT system (prospective phase). The comparison of automatic metric scores generated for output in this study and automatic metric scores for systems investigated in previous studies is not the main priority. However, since automatic metric scores are used as a means of easily judging the quality of MT output, we

calculated BLEU scores for the movie clips per corpus[106] so that MT researchers involved in automatic evaluation can rank our results in relation to other subtitle evaluation studies that use such scores. We also comment on the scores with reference to the qualitative data gathered during the current study, and discuss the possible meanings of the scores.

**Table 5.34:** Corpus AM: Automatic metric scores BLEU, NIST, METEOR, WER and PER

| Corpus A | | BLEU | NIST | METEOR | WER | PER |
|---|---|---|---|---|---|---|
| Clip 1 | 2007 | 25.26 | 3.61 | 48.98 | 59.13 | 51.07 |
| | 2008 | 51.18 | 4.64 | 65.58 | 37.63 | 35.48 |
| Clip 2 | 2007 | 13.07 | 3.0 4 | 39.05 | 60.16 | 52.84 |
| | 2008 | 32.99 | 3.84 | 55.67 | 49.59 | 42.27 |
| Clip 3 | 2007 | 33.58 | 3.84 | 52.03 | 49.18 | 45.90 |
| | 2008 | 50.06 | 4.54 | 64.63 | 38.25 | 35.51 |
| Clip 4 | 2007 | 22.58 | 3.65 | 47.00 | 50.54 | 46.15 |
| | 2008 | 45.70 | 4.64 | 63.93 | 35.16 | 32.41 |
| Clip 5 | 2007 | 34.30 | 3.51 | 49.79 | 45.37 | 36.97 |
| | 2008 | 32.62 | 3.12 | 46.32 | 51.26 | 43.69 |
| Clip 6 | 2007 | 20.49 | 3.75 | 51.66 | 53.33 | 43.70 |
| | 2008 | 46.76 | 4.73 | 68.43 | 36.29 | 28.14 |

[106] Using the MaTrEx system we were able to generate five different automatic metric scores and these are presented in Tables 5.34-5.36. However, in our discussion we mention only BLEU scores for reasons of comparison, but the reader is referred to Doddington (2002), Banerjee & Lavie (2005), Niessen et al. (2000) and Leusch et al. (2003) for discussions on the other scores presented. We calculated the scores in November 2007 when we were conducting the evaluation sessions (given on the first line of Tables 5.34-5.36). These scores are for the subtitles we used in all the end-user evaluation sessions, and the qualitative data gathered are discussed in relation to these scores only. We generated a second set of BLEU scores in June 2008, after the MaTrEx system had been modified and updated (given in the second line of Tables 5.34-5.36), but we do not have any qualitative data to support these scores.

**Table 5.35:** Corpus BM: Automatic metric scores BLEU, NIST, METEOR, WER and PER

| Corpus B | | BLEU | NIST | METEOR | WER | PER |
|---|---|---|---|---|---|---|
| Clip 1 | 2007 | 25.59 | 3.75 | 51.26 | 57.52 | 48.92 |
| | 2008 | 52.57 | 4.81 | 67.39 | 36.02 | 33.33 |
| Clip 2 | 2007 | 14.50 | 3.27 | 43.54 | 59.34 | 52.84 |
| | 2008 | 32.57 | 3.83 | 55.55 | 51.21 | 42.27 |
| Clip 3 | 2007 | 32.44 | 3.89 | 52.65 | 49.72 | 46.44 |
| | 2008 | 51.15 | 4.66 | 66.10 | 36.06 | 34.42 |
| Clip 4 | 2007 | 24.04 | 3.76 | 49.92 | 47.28 | 44.56 |
| | 2008 | 50.05 | 4.83 | 66.35 | 32.41 | 29.67 |
| Clip 5 | 2007 | 36.17 | 3.66 | 48.68 | 44.53 | 35.29 |
| | 2008 | 31.99 | 3.04 | 42.47 | 51.26 | 44.53 |
| Clip 6 | 2007 | 18.48 | 3.44 | 45.79 | 56.29 | 47.40 |
| | 2008 | 46.76 | 4.73 | 68.43 | 37.03 | 28.14 |

**Table 5.36:** Corpus CM: Automatic metric scores BLEU, NIST, METEOR, WER and PER

| Corpus C | | BLEU | NIST | METEOR | WER | PER |
|---|---|---|---|---|---|---|
| Clip 1 | 2007 | 25.71 | 3.57 | 47.17 | 60.21 | 53.76 |
| | 2008 | 56.64 | 5.02 | 70.02 | 33.33 | 30.10 |
| Clip 2 | 2007 | 15.17 | 3.31 | 44.43 | 60.16 | 52.03 |
| | 2008 | 43.42 | 4.20 | 63.91 | 43.90 | 37.39 |
| Clip 3 | 2007 | 28.77 | 3.71 | 48.60 | 50.27 | 46.44 |
| | 2008 | 48.22 | 4.51 | 63.48 | 38.25 | 36.61 |
| Clip 4 | 2007 | 24.99 | 3.84 | 49.29 | 49.45 | 43.95 |
| | 2008 | 58.07 | 4.99 | 68.70 | 29.12 | 27.47 |
| Clip 5 | 2007 | 35.71 | 3.69 | 51.27 | 46.61 | 37.28 |
| | 2008 | 32.94 | 3.12 | 42.33 | 51.26 | 44.53 |
| Clip 6 | 2007 | 22.80 | 3.69 | 48.59 | 53.33 | 44.44 |
| | 2008 | 42.57 | 4.54 | 66.72 | 41.48 | 30.37 |

Looking at the BLEU scores per clip (See first calculation, Table 5.37 below), we can see that in all three corpora, Clip 2 scored the lowest (13.07, 14.50, 15.17) and Clip 5 scored the highest (34.30, 36.17, 35.71).

**Table 5.37:** BLEU scores for the six movie clips per corpus

| BLEU | | Corpus AM | Corpus BM | Corpus CM |
|---|---|---|---|---|
| Clip 1 | 2007 | 25.26 | 25.59 | 25.71 |
| | 2008 | 51.18 | 52.57 | 56.64 |
| Clip 2 | 2007 | **13.07** | **14.50** | **15.17** |
| | 2008 | **32.99** | **32.57** | **43.42** |
| Clip 3 | 2007 | 33.58 | 32.44 | 28.77 |
| | 2008 | 50.06 | 51.15 | 48.22 |
| Clip 4 | 2007 | 22.58 | 24.04 | 24.99 |
| | 2008 | 45.70 | 50.05 | 58.07 |
| Clip 5 | 2007 | **34.30** | **36.17** | **35.71** |
| | 2008 | **32.62** | **31.99** | **32.94** |
| Clip 6 | 2007 | 20.49 | 18.48 | 22.80 |
| | 2008 | 46.76 | 46.76 | 42.57 |

This situation changed for the 2008 scores, with Clip 5 scoring the lowest in each corpus (32.62, 31.99, 32.94) and Clip 2 showing a dramatic increase, particularly in the case of Corpus CM (32.99, 32.57, 43.42). However, when we calculated the BLEU scores for the entire set of movie clips, we obtained the results presented in Table 5.38, which show similar scores across the three corpora. The 2008 scores show an increase for all three corpora.

**Table 5.38:** BLEU scores for the three corpora for all six clips (calculated in 2007 and 2008)

| | Corpus AM | Corpus BM | Corpus CM |
|---|---|---|---|
| All 6 clips: 2007 | 25.97 | 26.29 | 26.11 |
| All 6 clips: 2008 | 45.35 | 46.44 | 49.29 |

The 2007 scores in both tables are relatively low in comparison with Volk (2008), but are an improvement on Armstrong (2007) and Koehn (2005). The 2008 scores in both tables are closer to the kind of results Volk (ibid) reported. We can see from Table 5.37 that the BLEU scores for all clips except for Clip 5 improved after changes were made

to the EBMT system, in some cases by over 100%. From the qualitative data we analysed, it is reassuring that even though our BLEU scores obtained are considered 'low', 49% of Corpus AM subjects, 14% of Corpus BM subjects and 45% of Corpus CM subjects would definitely or perhaps consider using the subtitles as they were presented in this study (cf. Table 5.32). We must bear in mind that these subtitles are raw MT output, and the scores obtained were calculated using human reference translations, a comparison which has since been shown to misrepresent the true value of the MT quality (Volk & Harder 2007). Normally the BLEU scores we obtained for our data would be disregarded without much consideration. However, the feedback from the subtitle viewers contradicts the supposed 'bad quality' of the subtitles, and highlights the many positive factors that are kept hidden by automatic metrics, for example relating to style and acceptability. Focusing in particular on the results for Clips 2 and 5, 36% of subjects ranked Clip 5 higher than Clip 2 for comprehension and 34% of subjects ranked them in the opposite order, with 30% of subjects ranking them the same. This is a small example to show that even though the BLEU scores for these two clips differed by 20.6 (AM) points, 21.67 (BM) points and 20.01 (CM) points depending on the corpus, (with Clip 5 scoring the highest for each corpus), the quantitative data collected during the interview questionnaire tells a slightly different story.

Looking at the results in Table 5.39 overleaf, Corpus BM obtained the highest BLEU score by 0.18 over Corpus CM and by 0.32 over Corpus AM. In both cases the margin is too small to rank the corpora in terms of 'better quality' output. The same applies to the average mean scores of scale variables for Corpus BM, which is slightly higher in each category. However, given that the scores for each of the corpora are very close, we are unable to distinguish between the 'best' and 'worst' corpus.

**Table 5.39:** BLEU scores and mean scores for the four quality characteristics for each of the corpora

| | BLEU | Comprehension | Style | Error | Overall Satisfaction |
|---|---|---|---|---|---|
| **Corpus AM** | 25.97 | 3.10 | 3.35 | 2.63 | 2.30 |
| **Corpus BM** | 26.29 | 3.35 | 3.87 | 2.88 | 2.60 |
| **Corpus CM** | 26.11 | 3.17 | 3.61 | 2.82 | 2.85 |

The results highlight no significant changes in scores, even though the independent variables (corpus size, source language repetitions, corpus homogeneity) differ between the corpora. Despite the results being reversed for the 2008 scores, with an almost 3 point difference between Corpus CM and BM (Table 5.38), the recorded increases were very minimal. It is certainly encouraging to see the significant increase in the second set of BLEU scores. Qualitative and quantitative results already showed a link between somewhat acceptable machine-generated subtitles and relatively low BLEU scores. We could assume that the findings from a qualitative analysis using the new set of subtitles (2008) would be an improvement on the findings for the current study. We could also follow in the footsteps of Volk (ibid) and calculate a new set of BLEU scores using post-edited output as the reference translation. It was also suggested by some of the subjects that the MT output (as presented during the evaluation sessions) be used as material for post-editors.

There is no doubt that the use of automatic metrics is an extremely cheap and fast method of obtaining a quality benchmark for MT output. Nonetheless when machine-translated output is used in a domain such as AVT, the user is to a certain extent reliant on the translations for understanding and enjoyment purposes (cf. Gottlieb's claim that approximately 32% of the semantic load is communicated to the target audience through writing on the screen), especially in cases where the viewer does not understand the soundtrack language. In addition text is used in conjunction with an image and sound, in contrast to simply reading and understanding a text, which makes it essential that a human evaluation is conducted. This is in line with the importance of reception studies expressed by the AVT community. When this is carried out in conjunction with

the use of automatic metrics, we can verify the 'meaning' of the scores in this context. We can see from our results above that low BLEU scores do not necessarily mean unacceptable subtitles. There are many factors that determine the acceptability of subtitles, and a broader approach than that associated with the sole use of automatic metrics is needed.

## 5.3 Summarising Results

During the retrospective phase we conducted a quantitative and qualitative analysis of the data gathered during the evaluation sessions. We should refer here to the inter-subject agreement values (see section 5.1.7) which deemed Corpus BM as *substantial* agreement and Corpus AM and Corpus CM as *fair* agreement, following an analysis of the three continuous variables of comprehensibility, errors and style.

### Comprehensibility

None of the statistical tests returned significant results between the corpora for questions related to comprehensibility. In addition there were no qualitative questions included in the questionnaire specifically to measure comprehensibility. However, we noted earlier that often subjects' responses to questions from another category could also apply to this category. Within the *style* category four subtitles were highlighted for being incomprehensible, three of which were generated by Corpus AM. Within the *well-formedness* category comprehension issues was a sub-category in each corpus. Corpus CM had the highest number of codes in this sub-category (21) followed by Corpus AM (12) and Corpus BM (10). Despite there being no significant inter-corpus difference in comprehensibility scores, Corpus BM ranked highest (3.35).

We cannot rank the corpora in terms of comprehensibility based on these subjects' responses. All corpora received negative comments regarding the comprehensibility of the subtitles. However, due to the *substantial* inter-subject agreement of Corpus BM, coupled with the fact the comprehensibility score was higher than the other two corpora, we can deem Corpus BM subtitles as more comprehensible than either Corpus AM or Corpus CM.

### Readability

We saw earlier that subjects judged Corpus CM as having the lowest number of subtitles out of context and the speed of the subtitles was the most suitable. Therefore the readability of Corpus CM subtitles is ranked higher than the other two corpora.

### Style

From the quantitative and qualitative results we found that Corpus AM subtitles were deemed to be written in the least appropriate style. The low style ratings, coupled with the earlier finding of Corpus AM subjects rating the errors as most annoying (thus reducing the well-formedness of the subtitles) and their comments on the style of subtitles as something that negatively amused them, suggests that Corpus AM subtitles are less acceptable than the other corpora. The style of Corpus BM and Corpus CM subtitles is deemed more appropriate than Corpus AM. Looking at the mean scores for style, Corpus BM is ranked higher (3.87) than Corpus CM (3.61). This result is supported further by the inter-subject agreement value for Corpus BM. Therefore we could say the style of Corpus BM subtitles is the most appropriate.

### Well-formedness

Corpus CM subjects noted a significantly higher number of C1 errors than the other two corpora, and reported a lower number of well-translated subtitles, which together reduce the well-formedness of the subtitles. However Corpus AM subjects ranked the observed errors as most annoying of the three corpora. Even though the difference in mean scores for errors was not significant, Corpus BM subtitles were deemed the least annoying. Once again, we must also consider the *substantial* inter-subject agreement for Corpus BM. These results rank Corpus BM subtitles as the most well-formed.

### Overall

The mean scores for satisfaction suggest subjects were most satisfied with Corpus CM subtitles. When asked to comment on their satisfaction or dissatisfaction, Corpus CM received the highest number of positive comments and the lowest number of negative comments. Corpus AM received the highest number of positive comments on whether the subject would use the subtitles if they did not understand the soundtrack. These findings suggest that Corpus CM subtitles are the most satisfactory. However, given that the inter-subject agreement for all three corpora was deemed as *fair* and the qualitative data gathered on subjects' satisfaction consisted of additional comments

rather than answers to structured questions, the findings for satisfaction are not strong enough to lend support to the findings on intelligibility or acceptability for any of the corpora. The main findings of the *overall* category answer questions relating to subtitle repetition and the use of EBMT subtitles for watching movies on DVD.

### Relating the Findings to the Independent Variables

We now relate these findings to our independent variables, corpus size, number of SL repetitions and homogeneity. Findings from Corpus CM show that readability of machine-generated subtitles is improved if we increase the size of the corpus and with that the number of SL repetitions, and the corpus heterogeneity. Corpus CM was also the corpus with the highest number of alternative translations deemed acceptable by the evaluators in the given context. If increasing the number of TL translation segments increases the readability of the MT output, we cannot say the same for comprehensibility and acceptability. Corpus BM generated subtitles that were the most comprehensible, exhibited the most appropriate style and were the most well-formed. Therefore Corpus BM subtitles are the most comprehensible and acceptable. During the prospective phase Corpus BM was judged by the three evaluators as having generated the highest number of acceptable EBMT translations and it received the highest BLEU score (albeit very marginally in both instances) during the retrospective phase. On the one hand, the finding from the retrospective phase that Corpus BM EBMT subtitles were deemed the most comprehensible contradicts the assumption that corpus size and the number of SL repetitions improve intelligibility and acceptability of machine-generated subtitles. On the other hand, this same finding supports the results of the human evaluation of EBMT subtitles in a non-AVT environment (albeit one in which rich contextual information was available), and the BLEU scores.

The BLEU scores generated for the three corpora would normally be thought of as low scores (mid-twenties) and the EBMT output would possibly be discarded as 'bad quality' output. However, an analysis of the qualitative data showed that even low scoring subtitles can be considered 'acceptable' in particular contexts, and highlighted the importance of the human evaluation sessions when dealing with machine-translated subtitles.

We now relate the findings to the additional independent variables, namely whether or not subjects were listening to a known or unknown language soundtrack, subjects'

linguistic background and subjects' prior knowledge. The findings showed that subjects who have knowledge of the soundtrack language perceive the subtitles as more comprehensible (intelligibility) and more well-formed (acceptability). There is no difference for readability (intelligibility) or for style (acceptability).

Subjects with no formal language training judged the subtitles to be more comprehensible than subjects with formal language training (intelligibility). There was no difference for well-formedness or style (acceptability).

The findings showed that Corpus AM subjects with PK deemed subtitle errors as the most annoying (well-formedness), and the style of the subtitles was the least appropriate (well-formedness). These findings correspond to the previous findings (cf. Tables 5.6 and 5.8) when we did not consider PK as an independent variable. In relation to readability, Corpus AM subjects with PK deemed the subtitles as less readable than Corpus AM subjects with no PK. Corpus CM subjects with PK deemed the subtitles more satisfactory than Corpus AM subjects with PK. And Corpus CM subjects with PK deemed the subtitles more satisfactory than Corpus CM subjects with no PK.

When we conducted the quality check of the subtitles used in the training corpora, the format of the questionnaire used to gather the responses was the same at that used in the retrospective evaluation sessions. This means that we can compare the judgements of subjects looking at human subtitles provided on purchased DVDs and the EBMT-generated subtitles. Table 5.40 below shows the mean scores for comprehensibility, errors, style and overall satisfaction for human-generated subtitles and machine-generated subtitles (per corpus) as judged by human evaluators:

**Table 5.40:** Quality characteristic mean scores for human and machine-generated subtitles

| Subtitles | Comprehensibility | Errors | Style | Overall Satisfaction |
|---|---|---|---|---|
| **Human** | 5.33 | 4.45 | 4.60 | 5.03 |
| **Corpus AM** | 3.10 | 2.63 | 3.35 | 2.63 |
| **Corpus BM** | 3.35 | 2.88 | 3.87 | 2.78 |
| **Corpus CM** | 3.17 | 2.82 | 3.61 | 2.69 |

The judgement scores for human-generated subtitles are clearly higher than those obtained during the retrospective evaluation sessions. This result was to be expected given that human subtitles are proofed before being included on a DVD and the machine-generated subtitles are raw EBMT output. However, we must note that the human-generated subtitles also received a certain amount of criticism and they did not receive top scores for any of the four rating scales, even though they would be considered the 'gold standard' in terms of automatic evaluation methods. This result validates the carrying out of human evaluation of machine-generated subtitles. However, in future research, smaller studies could be adopted, such as those conducted at the shared tasks, but within an AVT environment. The results also provide a kind of benchmark for each of the quality characteristics which future automated studies could use.

The findings presented here indicate the need for a core homogeneous corpus which is supplemented by more heterogeneous corpus. This supplementary corpus can be used to increase the corpus size and the number of SL repetitions. This will have a positive effect on the intelligibility and acceptability of machine-generated subtitles. However, the extent of this effect will need to be investigated further.

## 5.4 Concluding Remarks

This chapter presented and discussed the data collected during the retrospective phase of this study. The chapter was split into a quantitative and qualitative analysis. Both sections presented, analysed and discussed the results in detail with reference to the research questions throughout. Following the retrospective phase, we introduced automatic metrics commonly used to evaluate MT output in the research community,

therefore relating this study to previous and current MT evaluation research. The chapter concluded with a summary of the results. In the next chapter we will discuss these results in relation to the aims of the study overall. We will look at the methodology and suggest possible improvements, and we will mention future avenues of research.

# - Chapter 6
# - Discussion and Conclusions

# 6 Discussion and Conclusions

## 6.1 Aims of the Study

The aims of this study were to establish whether target language subtitles produced by an EBMT system are considered intelligible and acceptable by viewers of movies on DVD (RQ1), and whether a relationship exists between the 'profiles' of corpora used to train an EBMT system, on the one hand, and viewers' judgements of the intelligibility and acceptability of the subtitles produced by the system, on the other (RQ2). These research questions were primarily investigated through human evaluation of MT output. Human evaluation of MT has been conducted since research into developing MT systems began, but there are certain disadvantages associated with this kind of evaluation, including costs, time and the recruitment of suitable subjects. Since the introduction in 2002 of automatic metrics, evaluation techniques have tended to move away from human to automatic. That said, as we outlined in the Introduction, the idea of introducing automated techniques into the subtitling process is relatively new. Subtitles are texts that are situated among three additional semiotic channels, namely image, sound effects and speech. Therefore, we argue that the evaluation of this text type is different to the evaluation of traditional texts. It was unknown whether text-based metrics, such as the commonly used BLEU, would generate scores that would reflect the 'true value' of the subtitle quality. In addition, there were no previous in-depth human evaluation studies that we could use as our evaluation model.

Therefore in order to meet the aims proposed we developed a human evaluation model to conduct end-user evaluations of subtitles (retrospective phase), combined with a corpus-analysis phase (prospective phase). During the corpus-analysis phase we created corpus profiles of our three training corpora Corpus AM, Corpus BM and Corpus CM, noting the corpus size and the number of SL repetitions (within the corpora and between the test data and the corpora), and we decreased the homogeneity of the corpora as the two other factors increased. Individual human evaluation sessions were conducted with 44 native-German speakers to gather end-user judgements on the intelligibility and acceptability of EBMT-generated subtitles. From these data we could establish whether a relationship existed between the profiles of the corpora used to train the system (independent variables) and viewers' judgements on intelligibility and acceptability

(dependent variables). In addition we generated automatic metric scores for the three training corpora in order to situate this research among current MT evaluation studies, and to investigate the relationship between these metrics and human judgements.

## 6.2 Findings of the Study

To answer RQ1 the data from the retrospective phase showed that the subtitles generated by all three training corpora were deemed intelligible and acceptable to a certain degree. Subjects were asked if they would use subtitles of a similar standard to the subtitles shown on the clips with the unknown language soundtrack. The findings were based on the three movie clips with the unknown language soundtrack and they showed that in 38% of Corpus AM cases, 20% of Corpus BM cases and 40% of Corpus CM cases, end-users said they would use this standard of subtitles. We should also comment that inter-subject agreement for Corpus BM was considered *substantial*, while the other two corpora were deemed *fair* (based on scale scores).

In response to RQ2 we examine the data from both phases. The data from the prospective phase showed that an increase in corpus size (and resultant decrease in homogeneity) was accompanied by an increase in subtitle repetition (segment and sub-segment levels). Thus Corpus CM, the largest and least homogeneous training corpus used, exhibited the highest levels of repetition at segment and sub-segment level. When three evaluators were asked to rate the movie clips subtitled using an EBMT system trained on our three corpora, however, they deemed Corpus BM to have produced the most acceptable subtitles in the given context. This result shows that higher levels of repetition in the training corpus do not necessarily mean higher acceptability of subtitles. This evaluation was however conducted in a non-AVT environment, with evaluators reading subtitles on paper.

Following this we investigated whether alternative translations of subtitles in the movie clips located in the corpora, but not chosen by the EBMT system, could be used in the context of *Harry Potter* movies. Corpus CM contained the highest number of alternative translations deemed acceptable by the three evaluators. That said a large proportion of these were possibly accounted for by Corpus AM, a finding which highlights the importance of homogeneous training data. The investigation of "alternative translations" thus showed that increasing the corpus size and number of SL

repetitions (while decreasing the homogeneity of the training corpus) generated more alternative translations that could be re-used in the *Harry Potter* context.

The retrospective phase data show only non-significant differences for comprehensibility between the three training corpora (with Corpus BM exhibiting the highest mean score); the readability of Corpus CM subtitles is higher than the other two corpora; the well-formedness of Corpus BM subtitles is higher than the other two corpora; and the style of Corpus BM and Corpus CM subtitles is more appropriate than the style of Corpus AM subtitles. We are unable to distinguish statistically between Corpus BM and Corpus CM for comprehensibility and the appropriateness of style. However, we mentioned previously that the mean scores for Corpus BM were higher for both characteristics, and the inter-subject agreement value was deemed *substantial*. This suggests Corpus BM subtitles are more comprehensible and written in a more appropriate style than those based on the other two corpora. This suggests that the percentage of subjects who would use subtitles of a similar standard is more reliable for Corpus BM than for the other two corpora.

From the findings:

- We are unable to rank the corpora in terms of intelligibility. Corpus BM subtitles are deemed the most comprehensible and Corpus CM subtitles are deemed the most readable
- The well-formedness of Corpus BM subtitles is deemed the best of the three corpora, and combining this with style, the acceptability of Corpus BM subtitles is higher than either Corpus AM or Corpus CM subtitles

Previous research in the MT community showed that an increase in corpus size resulted in an increase in the quality of the MT output.[107] Looking at the quantitative data from the prospective phase we would assume that Corpus CM subtitles would be considered the most intelligible and most acceptable. The qualitative data, on the other hand, showed a different result. Subjects judged Corpus BM as having generated the most acceptable EBMT subtitles in the given context. This result undermines the assumption

[107] We acknowledge that research has shown that there seems to be a point beyond which adding further examples does not improve the overall translation quality (Mima et al. 1998). However, the corpora used in this study are relatively small compared to others in the EBMT literature, and therefore we were unlikely to witness this kind of diminishing returns effect.

that by maximising training corpus size, the number of SL repetitions, and the number of alternative translations contained in the training corpus we also maximise acceptability of MT output.

Based on human judgements for comprehensibility, readability, style and well-formedness, the retrospective phase showed that Corpus CM generated the most readable subtitles and Corpus BM generated the most comprehensible and acceptable subtitles. The readability result, on the one hand, supports the assumption that an increase in corpus size and SL repetitions, and an increase in acceptable TL translations means higher quality MT output. However, this assumption does not hold for comprehensibility and acceptability, on the other. The qualitative data relating to acceptability from the prospective phase is consistent with the acceptability result in the retrospective phase. When corpora were ranked according to BLEU scores, Corpus BM came first, followed by Corpus CM and then Corpus AM (albeit by a very small number of points). These results agree with our preliminary assumptions made during the prospective phase regarding acceptability following the human evaluation. The research also found that relatively low BLEU scores did not correlate with human judgements of intelligibility and acceptability.

In summary, there is a relationship between the corpus profiles and our dependent variables. Corpus CM output was deemed more readable than Corpus BM output. This shows a relationship between increasing corpus size and increasing readability. Corpus BM output was deemed more comprehensible and acceptable than Corpus AM output. The same relationship does not hold true for comprehensibility and acceptability. During the prospective phase we noted an increase in corpus size, in the number of SL repetitions and a decrease in homogeneity resulted in an increase in the number of alternative translations in the corpus deemed acceptable by human evaluators. We thought that this might have been a factor that would improve the intelligibility and acceptability during the retrospective phase. However, this assumption did not hold true. Considering the contribution from each of the 'sub-corpora', the number of alternative translations relative to the size of Corpus BM and Corpus CM represented low 'added value'. This finding would have to be investigated further, as the EBMT system used in this study implements a similarity metric, and therefore would not offer a kind of ambiguity in the system (Somers 1999:92).

In addition to addressing RQ1 and RQ2 above we aimed to draw some conclusions about three related subsidiary research questions:

**RQ3**: If the viewer understands the soundtrack, are they more accepting or less accepting of the EBMT subtitles?

An analysis of the quantitative and qualitative data showed that subjects were more accepting of EBMT subtitles when they understood the soundtrack. There was a significant intra-corpus difference between the comprehensibility scores depending on the soundtrack language (small effect size - see Table 5.10). There was a similar result for errors, as errors noted in clips with the Dutch language soundtrack were rated as more annoying than the errors noted in the English language soundtrack (see Table 5.11). There is no significant difference between the scores for style when we take the soundtrack language into consideration. However, the style scores are all lower for the Dutch language clips. The data suggest that if viewers have knowledge of the soundtrack language they perceive the subtitles as more comprehensible (comprehensibility scale) and more well-formed (error scale).

We also discussed one categorical variable related to comprehensibility that was significantly different depending on the soundtrack language. We found that Corpus BM viewers sometimes used the soundtrack to understand the movie clip rather than reading the subtitles. This result is perhaps inevitable given that listening is easier than reading (Koolstra et al. 2002). This phenomenon does not, however, mean that subtitles provided on the movie clips with the English language soundtrack are deemed less comprehensible, as our research has shown. Table 5.29 shows that subjects who noticed differences between the quality of the subtitles, depending on the soundtrack language, were more satisfied with the subtitles when they were listening to a known language soundtrack. Previous studies and received wisdom may lead one to the assumption that if a viewer knew the soundtrack language that they would be more critical of the subtitles, because they would have a better basis on which to judge the subtitles (cf. Armstrong et al. 2006c). The findings of the current study show this assumption to be unfounded. A critic might say, however, that viewers are less likely to be critical of the subtitles if they know the soundtrack language, precisely because they do not have to

rely on the subtitles to understand the movie. The availability of another semiotic channel means that the quality of the subtitles is less critical.

**RQ4**: If the viewer has a 'linguistic background' (LB), are they more accepting or less accepting of the EBMT subtitles?

Linguistic background was a factor we thought might influence viewers' judgements on the intelligibility and acceptability of EBMT-generated subtitles. The quantitative data showed that subjects with formal language training deemed the subtitles to be less comprehensible (continuous variable). The lower scores might reflect a lower tolerance of grammatical errors among subjects with a more formal training in language. Linguistic background had no significant effect on how subjects rated errors or style. When we investigated the effect of LB on the categorical variables, the findings showed that judgements on the readability (Corpus AM) of the subtitles and the comprehensibility (Corpus CM) of the subtitles from subjects with a linguistic background did not improve the results in any way.

**RQ5:** If the viewer has prior knowledge (PK) of the movie or related material such as books, are they more accepting or less accepting of the EBMT subtitles?

The quantitative data showed no significant difference for comprehensibility scores. For the other three continuous variables (errors, style and overall satisfaction), there were five significant results that indicated an interaction effect between the corpus and prior knowledge. The four results with medium and large effect size are discussed here:

- Satisfaction scores were higher when subjects had PK within Corpus CM (large)
- Satisfaction scores were higher for Corpus CM than for Corpus AM when subjects had PK (large)
- Style scores were lower for Corpus AM than the other two corpora when the subjects had PK (medium)
- Errors scores were lower (i.e. errors were more annoying) for Corpus AM than Corpus BM when the subjects had no PK (medium)

The results for the continuous variables suggest that Corpus CM subjects with PK deemed the style of the subtitles more appropriate than Corpus AM subjects with PK.

The inter-corpus mean scores for errors and style were lower for Corpus AM than the other two corpora when we did not take PK into account (cf. Tables 5.6 and 5.8). When we considered the impact of PK on subjects' judgements, once again the scores for Corpus AM subjects were lower than those of the other two corpora. These findings suggest that PK does not have an effect on the low scores given by Corpus AM subjects.

Three categorical variables also returned statistically significant results when PK was taken into account, and all of these results related to Corpus AM. The findings suggest that Corpus AM subjects with no PK are more accepting of the subtitles.

When analysing the data set, we made some additional observations regarding viewer judgements of machine-generated subtitles. These observations can be categorised as (1) subtitles that polarise subjects, (2) subtitles that are not recognised, and therefore perceived as incorrect, and (3) subtitles that elicit low scores, but are nonetheless acceptable, and they allow us to problematise our work.

Firstly, in this study we acknowledge the subjective nature of human evaluation, and therefore it was not surprising that we observed subjects who were unable to agree on particular subtitles. Secondly, we also observed subjects who rated the MT subtitles as incorrect, because they did not recognise *Harry Potter*-specific terminology. It was interesting to note that many of these subjects had prior knowledge of the data set. Lastly, the third observation emphasises the need for a human evaluation study. On some occasions subjects gave low judgements on the intelligibility and acceptability of the subtitles, but said that they would use the subtitles in certain contexts. Viewers' opinions such as these are important for evaluating subtitles, and these judgements would not be gathered through automatic methods alone.

## 6.3 Limitations of the Research

In relation to corpus profiling during the prospective phase, we experienced some problems when we used the SDL Trados Analyse Tool. Errors occurred when we generated statistics with the files containing Corpus BM and Corpus CM, as the file size was too large. However, this corpus is extremely small compared to other training corpora used in Corpus-Based MT. In addition, the tool was not able to generate all of

the profiling data relating to the repetitiveness of the subtitles. We could not tell from the data on repeated source language subtitles whether the subtitle had exactly one German translation or whether the subtitle had two or more different German translations. This investigation had to be conducted manually, which is quite time-consuming. However, this information is useful in advance of training the system, and would give some insight into the expected quality of the MT output. Therefore if we were to once again create corpus profiles, it would be advisable to use a script-based Unix program (e.g., Perl scripts). In addition, Microsoft Word was a suitable text processing tool for the demands of this study given the relatively small size of the corpora. If, however we were dealing with much larger corpora, an alternative method would have to be used to locate TL subtitles, for example. It would also be beneficial if we could avoid aligning the subtitles using the method outlined here, and follow a similar approach to Volk (2008). Volk's approach is somewhat dependent on the language pair, and the approach favours closely related languages (e.g. Swedish, Danish and Norwegian). In the current study English and German are related, but we experienced difficulties when aligning subtitles at sub-segment level written in the German compound past tenses.

MT system development is an ongoing process, and improvements are integrated into the systems with the aim of improving the quality of the output. Therefore a study such as the current one uses MT output during the evaluation process which could be considered 'out-of-date' as soon as it is generated. In Chapter 5 we presented BLEU scores for the two sets of EBMT output; the first set of subtitles was used in the current study's evaluation sessions, and the second set was not evaluated by human judges. The second set of BLEU scores showed a significant increase over the first. This would suggest that the human judgements of this second set of subtitles would also show an improvement on the findings presented for the current research. These BLEU scores are purely an indication of what we might expect if we conducted a second human evaluation, but improved human judgements are, of course, not guaranteed.

Following an analysis of the data collected during the retrospective phase, possible weaknesses with the interview questionnaire were identified. Firstly, the representation of scales in the questionnaire was not implemented in the same way as those used in other human MT evaluations (e.g. Pierce et al. 1966, Van Slype 1979, ACL 2007,

2008). We asked subjects to rate the comprehension, errors and style on a scale of 1-6, and simply explained that 1 was the lowest score (e.g. least comprehensible) and 6 was the best score (e.g. most comprehensible). In a number of other human MT evaluations each number on the scale is assigned a specific explanation. Perhaps this design would explain more clearly the rating assigned to a clip by viewers, and the data could be compared to other studies that use these types of scales. That said the design of our scales follow a similar model presented in Elliot et al. (2004) and Babych et al. (2005).

Secondly, the layout of the questionnaire could have been more structured in advance of the evaluation sessions. We arranged the questions into categories representing the four quality characteristics after collecting the data. This meant that some of the data set overlapped between categories, and each category did not contain a balanced number of questions. We did not include any kind of scale within the readability category. This is something that could be revised for future studies.

Regarding the questionnaire design, some insight might have been gained from AVT reception study approaches, e.g. the kind of questions they ask to elicit judgements on the style of subtitles (even though these subtitles are human-generated as opposed to MT-generated). The questionnaires in the current study were designed from an MT evaluation perspective.

## 6.4 Contribution to the Literature

The current study contributes to the literature in the following ways:

- It represents the first large-scale human evaluation of automatically translated movie subtitles: previous studies used only one evaluator (Popowich et al. 2000) or at the most six evaluators (Armstrong et al. 2006c)
- The study is also the first end-user evaluation of machine-translated subtitles in an audiovisual context. We have previously mentioned the differences between the evaluation of text-based material and audiovisual material. These differences motivate the evaluation of subtitles in a context where all the semiotic channels are available to the evaluators. We argued for the need for a context-based evaluation, and the use of real end-users of subtitles. The combination of these factors improves the ecological validity of the study
- The current study investigates questions other sources have ignored:

1. Within the EBMT literature exact matches have been described as exceptional matches rather than the rule, and therefore it seems common practice to consider exact matches as trivial and not worthy of further investigation. We argue that if we increase the number of exact matches, we increase the number of potential matches at sub-segment level. Therefore our approach aimed to identify the highest number of exact matches between the test and training data, and to test whether improvements could be observed due to an increase in exact matches
2. We took the topic of SL repetitions a step further when we also evaluated the reusability of the corresponding TL translations in a *Harry Potter* context. It is reported in the EBMT literature that the presence of overlapping examples could be problematic, where the examples are in conflict, on the one hand, or where systems do not use a similarity weighting metric, on the other. However, no studies have investigated the potential reusability of these 'alternative' translations before now. We showed that an increase in corpus size and in the number of SL repetitions in the training corpus produced the highest number of alternative translations deemed acceptable by evaluators. This result could lead us to assume this combination would produce the most intelligible and acceptable subtitles. That said, these alternative translations did not have a significant impact on the EBMT-generated subtitles, as two-thirds of Corpus CM-generated subtitles were the same as Corpus AM-generated subtitles. We found that in the case of subtitling, an increase in alternative TL translations improved only the readability of subtitles. Sets of alternative translations could also be thought of as helpful to subtitlers if EBMT is conceived of as an aid to human translators. However, this view is not shared by many subtitlers who are currently using SMT technology as an aid in the subtitling process (Martin Volk: personal communication)
3. We commented earlier in relation to viewers' judgements on subtitles when they were listening to a known and an unknown language soundtrack. Our findings are at variance with Armstrong et al.'s (2006c), and we acknowledge the possible counter argument to the findings. Either way this question needs to be addressed and given the conflicting results, no assumptions can yet be made
4. We asked subjects if they noticed any repeated subtitles during the clips. We wanted to investigate whether such repetitions, which could be generated by technology such as EBMT, distracted the subjects and therefore could be a

possible disadvantage of introducing such technology into the subtitling process. When asked this question only seven of the forty-four subjects reported noticing at least one repeated subtitle, and subjects did not comment on repeated subtitles during the course of the interview. Given that the current study dealt with only a small number of repeated subtitles, we are not able to generalise this finding. However, in this instance, we can argue that subjects' judgements on intelligibility and acceptability were not disrupted in any way by the presence of repetition. This point needs further investigation, but the finding is very relevant given the common arguments against an automated subtitling solution (cf. panel discussion Languages and the Media Conference 2006)

- Corpus profiling is an emerging area within MT research. Given that corpus-based approaches to machine translation dominate the MT field today, it is not surprising that MT developers want to know more about the content of the training corpus. Establishing corpus profiles allow researchers to derive correlations between the training corpus and the quality of the MT output (cf. Ozdowska & Way 2009)
- Lastly, the current study investigates factors that may have an effect on the evaluation of MT subtitle output but are ignored elsewhere

## 6.5 Future Research

In this study we focused on the intelligibility and acceptability of EBMT-generated subtitles from the end-user's perspective, and presented the subjects with raw EBMT output to evaluate. We know that when viewers use subtitles in a commercial setting (provided on DVD), they would be unlikely to watch subtitles of the quality presented here. However, data based on raw EBMT output represents a baseline for human evaluations, and any subsequent studies using post-edited output should achieve better results.

The motivation for introducing technology into the subtitling domain is to aid the subtitler. This means that machine-generated subtitles need some form of post-editing before they are included in AVT material. This point did arise during the evaluation process, as it was mentioned by some subjects as a possible 'next step' to increase intelligibility and acceptability. The discussion of post-editing subtitles is beyond the scope of the current study. That said it is an area worthy of further research. This could

involve a study of how automatically translating subtitles combined with post-editing could alleviate time pressures on subtitlers and reduce costs for subtitling companies. A different study could investigate the intelligibility and acceptability of post-edited subtitles using end-users, focusing on the overall benefits of generating subtitles in this way, with the aim of semi-automating the subtitling process on a larger scale.

In this study we opted for a large-scale human evaluation because it was lacking from the current literature, and we felt this type of evaluation would allow us to develop interview techniques and specific questions to elicit pertinent information. However, as mentioned earlier, we could gain a good insight from AVT reception studies when designing the questionnaire for future evaluations of machine-translated subtitles (e.g. Antonini 2005, Chiaro 2007). In addition, we acknowledge that large-scale human evaluations are time consuming and somewhat costly. Therefore, based on the evaluation model presented in this study, we would suggest future human evaluations could be conducted using an online evaluation model. The design of this model would be more succinct than the categories presented in our questionnaires, based on experiences reported on in this research. After analysing the data sets we are more aware of the kind of data we need to elicit, and the use of online methods could facilitate this requirement. Human evaluation conducted online gives the subjects flexibility in relation to completing the evaluation, reduces costs, and could attract large numbers of subjects. The online modules would allow different size studies to be conducted, depending on the complexity of the relationships being investigated, without much effort needed to customise each study. This approach would be very successful given that the ground work has been done already in this research, and the fact that multimodal texts such as subtitled movie clips are fully supported in an online environment. Just as the work of Armstrong et al. (2006c) served as an essential pilot study to the current research, the current study will provide the basis for a successful move from interactive to virtual human evaluation.

# – References

# References

**Ackroyd, Stephen and John A. Hughes.** (1992). *Data collection in context*. (2nd edition). Harlow: Longman Group (original work published in 1981).

**Aizawa, Teruaki, Terumasa Ehara, Noriyoshi Uratani, Hideki Tanaka, Naoto Kato, Sumio Nakase, Norikazu Aruga, and Takeo Matsuda.** (1990). A machine translation system for foreign news in satellite broadcasting. *IN: Proceedings of the 13th conference on Computational* Linguistics, Helsinki, Finland. pp. 308-310.

**Allen, Graham.** (2000). *Intertextuality*. London, USA and Canada: Routledge.

**Amigó, Enrique, Jesús Giménez, Julio Gonzalo and Luís Màrquez.** (2006). MT evaluation: human-like vs. human acceptable. *IN*: *Proceedings of the COLING/ACL main conference poster session*, Sydney, Australia. pp. 17-24.

**Antonini, Rachele.** (2005). The perception of subtitled humour in Italy. *Humor – International Journal of Humor Research*, 18 (2), pp. 209-225.

**Aparicio, Antonio, Michael Benis and Graham Cross.** (2001). Rates and Salaries Survey. ITI Bulletin. [Online]. Available from: <http://www.iti.org.uk/uploadedFiles/surveys/ITI2001R&S.pdf> [Accessed 10 March 2009].

**Armstrong, Stephen.** (2007). *Using EMBT to produce foreign language subtitles*. MSc Thesis. Dublin City University.

**Armstrong, Stephen, Marian Flanagan, Yvette Graham, Declan Groves, Bart Mellebeek, Sara Morrissey, Nicolas Stroppa and Andy Way.** (2006a). MaTrEx: Machine Translation Using Examples. *IN: TC-Star OpenLab Workshop on Speech Translation*, Trento, Italy.

**Armstrong, Stephen, Colm Caffrey, Marian Flanagan, Dorothy Kenny, Minako O'Hagan and Andy Way.** (2006b). Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation (EBMT). *IN: Proceedings of ASLIB Translation and the Computer Conference*. London. pp. 1-13.* (Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.)

**Armstrong, Stephen, Colm Caffrey, Marian Flanagan, Dorothy Kenny, Minako O'Hagan and Andy Way.** (2006c). Leading by Example: Automatic Translation of Subtitles via EBMT? *Perspectives*, 14 (3), pp.163-184.

**Arnold, Doug, Lorna Balkan, R. Lee Humphreys, Siety Meijer and Louisa Sadler.** (1994). *Machine Translation: An Introductory Guide*. Manchester: NCC Blackwell.

**Arnold, Doug, Sadler, Louisa, and R. Lee Humprheys.** (1993). Evaluation: An Assessment. *Machine Translation*, 8 (1-2), pp. 1-24.

**Austermühl, Frank.** (2001). *Electronic tools for translators.* Manchester: St. Jerome.

**Babych, Bogdan, Anthony Hartley and Debbie Elliott.** (2005). Estimating the predictive power of n-gram MT evaluation metrics across language and text types. *IN: Proceedings of the 10th Machine translation Summit (MT Summit X)*. Phuket, Thailand. pp. 412-418.

**Bailey, Kenneth D.** (1994). *Methods of Social Research*. 4th edition. New York/Oxford/Singapore/Sydney: The Free Press (a division of Macmillan, Inc.)

**Baker, Mona.** (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target,* 7 (2), pp. 223-43.

**Bakhtin, Mikhail.** (1981). *The dialogical imagination*. C. Emerson and M. Holquist (eds.) M. Holquist (trans.) Austin: University of Texas Press.

**Bakhtin, Mikhail**. (1986). *Speech genres and other late essays*. C. Emerson and M. Holquist (eds.) V. W. McGee (trans.) Austin: University of Texas Press.

**Banerjee, Satanjeev and Alon Lavie.** (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. *IN: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05),* Ann Arbor, Michigan. pp. 65-72.

**Barber, Theodore X.** (1976). *Pitfalls in human research: Ten pivotal points*. Elmsford, NY: Pergamon Press.

**Barthes, Roland.** (1977). Death of the Author *IN:* S. Heath (trans.) (ed.) *Image-Music-Text*. New York: Hill and Wang. pp. 142-148.

**Barzilay, Regina and Mirella Lapata.** (2005). Modeling local coherence: An entity-based approach. *IN: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, Michigan. pp. 141–148.

**Becquemont, Daniel**. (1996). Le sous-titrage cinématographique: contraintes, sens, servitudes. *IN :* Y. Gambier (ed.) *Les transferts linguistiques dans les médias audiovisuels*. Presses universitaires du Septentrion, Villeneuve d'Ascq. pp. 145–155.

**Bédard, Claude.** (2000). Mémoire de traduction cherche traducteur de phrases. *Traduire,* 186, pp. 41-49.

**Benis, Michael.** (1999). How the memory measured up. TransRef—The Translation Reference Centre. [Online]. Available from: <http://www.transref.org/default.asp?docsrc=/u-articles/Benis4.asp> [Accessed 10 March 2009].

**Benis, Michael**.(2000). Translation Memory from O to R. TransRef—The Translation Reference Centre. [Online].

Available from: <http://www.transref.org/default.asp?docsrc=/u-articles/Benis3.asp> [Accessed 10 March 2009].

**Bhatia, Vijay K.** (1993). *Analysing Genre*. London: Longman.

**Biber, Douglas, Susan Conrad and Randi Reppen.** (1998). *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

**Bowker, Lynne.** (2002). *Computer-aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

**Bowker, Lynne.** (2005). Productivity vs. quality? A pilot study on the impact of translation memory systems. *Localisation Focus*, 4 (1), pp.13–20.

**Bowker, Lynne and Melissa Ehgoetz**. (2007). Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation. *IN*: D. Kenny and K. Ryou (eds.) *Across Boundaries: International Perspectives on Translation Studies.* Newcastle, UK: Cambridge Scholars Publishing. pp. 209-224.

**Brøndsted, Katrine and Cay Dollerup.** (2004). The names in Harry Potter. *Perspectives: Studies in Translatology*, 12 (1), pp. 56-72.

**Cadwell, Patrick.** (2008). Readability and Controlled Language: Does the study of readability have merit in the field of controlled language, and is readability increased by applying controlled-language rules to texts? MA Thesis. Dublin City University.

**Callison-Burch, Chris, Miles Osborne and Philip Koehn.** (2006). Re-evaluating the role of BLEU in Machine Translation research. *IN: Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL).* pp. 249-256.

**Callison-Burch, Chris, Cameron Fordyce, Philip Koehn, Christof Monz and Josh Schroeder.** (2007). (Meta-) evaluation of Machine Translation. *IN: Proceedings of the Association for Computational Linguistics (ACL-2007) Workshop on Statistical Machine Translation*. pp. 136-158.

**Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder.** (2008). Further Meta-Evaluation of Machine Translation. *IN: Proceedings of the Association for Computational Linguistics (ACL-2008) Workshop on Statistical Machine Translation*. pp 70-106.

**Carbonell, Jaime and Yorick Wilks.** (1991). Machine Translation: An In-Depth Tutorial. *IN: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, University of California, Berkeley, USA. pp. ix.

**Carl, Michael and Andy Way.** (eds.) (2003). *Recent Advances in Example-Based Machine Translation*. Dordrecht/Boston/London: Kluwer Academic Publishers.

**Carroll, Mary.** (2004). Subtitling: Changing standards for new media? *LISA Newsletter Global Insider*, XIII, 3.5. [Online]. Available from: <http://www.lisa.org/globalizationinsider/2004/09/subtitling_chan.htm> [Accessed 10 August 2009].

**Cerón, Clara.** (2001). Punctuating subtitles: Typographical conventions and their evolution. *IN:* Y. Gambier and H.Gottlieb (eds). *(Multi)media Translation. Concepts, Practices, and Research*. Amsterdam/Philadelphia: John Benjamins Publishing. pp.173-188.

**Chandler, David**. (2002). Semiotics for Beginners. [Online]. Available from: <http://www.aber.ac.uk/media/Documents/S4B/semiotic.html> [Accessed 10 March 2009].

**Chandler, David**. (2007). *Semiotics: the basics*. 2nd edition. Oxford/New York: Routledge.

**Charmaz, Kathy.** (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage.

**Charniak, Eugene, Kevin Knight and Kenji Yamada.** (2003). Syntax-Based Language Models for Statistical Machine Translation. *IN: Proceedings of the Ninth Machine Translation Summit*. New Orleans: MT Summit. pp. 40–46.

**Chaume, Frederic.** (2004). Film Studies and Translation Studies: Two Disciplines at Stake in Audiovisual Translation. *Meta,* 49 (1), pp.12-24.

**Chesterman, Andrew and Emma Wagner.** (2002). *Can Theory Help Translators? A Dialogue Between the Ivory Tower and the Wordface.* Translation Theories Explained, (9). Manchester: St. Jerome.

**Chiaro, Delia.** (2006). How big should a sample be? [Online]. Available from: <http://www.est-translationstudies.org/Research%20issues/Samplesizechiaro.htm> [Accessed 10 March 2009].

**Chiaro, Delia.** (2007). The effect of translation on humour response: the case of dubbed comedy in Italy. *Doubts and Directions in Translation Studies*. Amsterdam/Philadelphia: John Benjamins Publishing. pp. 137-152.

**Choi, Jimmy.** (2006). One way or another something has to happen: Subtitling, Dubbing and the future of international film. [Online]. Available from: <http://ics.leeds.ac.uk/papers/vp01.cfm?outfit=ifilm&folder=17&paper=23> [Accessed 10 March 2009].

**Church, Kenneth and Eduard Hovy.** (1993). Good applications for crummy machine translation. *Machine Translation*, 8 (4), pp. 239-258.

**Circé, Karine.** (2005). *Traduction automatique, mémoire de traduction ou traduction humaine? Proposition d'une approche pour déterminer la meilleure méthode à adopter, selon le texte*. MA Thesis. University of Ottawa.

**Clough, Paul, Robert Gaizauskas, Scott S.L. Piao and Yorick Wilks**. (2002). METER: MEasuring TExt Reuse. *IN: Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-2002).* University of Pennsylvania, Philadelphia, USA: ACL-02. pp.152-159.

**Coffey, Amanda and Paul Atkinson.** (1996). *Making sense of qualitative data: complementary research strategies*. Thousand Oaks: Sage.

**Cohen, Jacob W.** (1988). *Statistical power analysis for the behavioral sciences*. 2nd edition. Lawrence Erlbaum Associates: Hillsdale, NJ.

**Copyright and Related Rights Act.** (2000). (Homepage). [Online]. Available from: <http://www.irishstatutebook.ie/2000/en/act/pub/0028/index.html> [Accessed 10 March 2009].

**Copyright, Designs and Patents Act.** (1988). (Homepage). [Online]. Available from: <http://www.opsi.gov.uk/acts/acts1988/UKpga_19880048_en_1.htm > [Accessed 10 March 2009].

**Coughlin, Deborah.** (2003). Correlating Automated and Human Assessments of Machine Translation Quality. *IN*: *Proceedings of MT Summit IX,* New Orleans, USA. pp. 63-70.

**Creswell, John.** (1998). *Qualitative inquiry & research design: choosing among five approaches*. London/Thousand Oaks: Sage.

**Creswell, John. W and Vicki L. Plano Clark.** (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks/London/New Delhi: Sage Publications.

**Crook, Mason N. and Harold P. Bishop.** (1965). *Evaluation of Machine Translation, Final Report.* Institute for Psychological Research, Tufts University, Medford, Mass.

**de Beaugrande, Robert and Wolfgang Dressler.** (1981). *Introduction to Text Linguistics*. Essex, UK: Longman.

**de Linde, Zoe and Neil Kay.** (1999). *Semiotics of Subtitling*. Manchester: St. Jerome.

**De Vaus, David.** (2002). *Surveys in Social Research*. 5th edition. London: Routledge.

**Delisle, Jean.** (2006). Criticizing Translations: The Notion of Disparity. *IN:* L. Bowker (ed.) *Text-based Studies: Lexicography, Terminology, Translation, In Honour of Ingrid Meyer*. Ottawa: University of Ottawa Press. pp.159-173.

**Denoual, Etiene.** (2005). The Influence of Example-data Homogeneity on EBMT Quality. *IN: Proceedings of the Second Workshop on Example-Based Machine Translation*, Phuket, Thailand. pp. 35-42.

**Díaz Cintas, Jorge.** (2005). Back to the Future in Subtitling. *IN: Proceedings of* the *Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation.* Saarbrücken, Germany: MuTra. pp 1-17.*

(Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.) [Online]. Available from: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_DiazCintas_Jorge.pdf> [Accessed 10 May 2009].

**Díaz Cintas, Jorge and Aline Remael.** (2007). *Audiovisual translation: subtitling*. Manchester, UK/Kinderhook, New York: St. Jerome.

**DiMarco, Chrysanne and Graeme Hirst.** (1990). Accounting for Style in Machine Translation. *IN: Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-1990)*, Austin, Texas, USA. pp. 148-153.

**DiMarco, Chrysanne.** (1994). Stylistic Choice in Machine Translation. *IN: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, Maryland, USA. pp. 32-39.

**Doddington, George.** (2002). Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. *IN: Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, USA. pp. 138-45.

**Dostert, Bozena H.** (1973). *Users' evaluation of machine translation*. Report RADC-TR-73-239, Rome Air Development Center, Griffiss Air Force Base, NY.

**Doyon, Jennifer B., Kathryn B. Taylor and John White.** (1999). Task-based Evaluation for Machine Translation. *IN: Proceedings of the Machine Translation Summit VII*. pp. 574-578.

**Du, Jinhua, Yifan He, Sergio Penkale and Andy Way.** (2009). MaTrEx: the DCU MT System for WMT 2009. *IN: Proceedings of the Fourth Workshop on Statistical Machine Translation, EACL 2009*, Athens, Greece (in press).

**Dunne, Ciarán.** (2008). *"We know them, but we don't know them": A Grounded Theory Approach to Exploring Host Students' Perspectives on Intercultural Contact in an Irish University*. PhD Thesis. Dublin City University.

**Dyson, Mary C. and Jean Hannah.** (1987). Towards a Methodology for the Evaluation of Machine-Assisted Translation Systems. *Computers and Translation*, 2 (3), pp. 163-176.

**Eco, Umberto.** (2004). *Mouse or rat? Translation as negotiation*. London: Phoenix.

**Elliot, Debbie, Eric Atwell and Tony Hartley.** (2004). Compiling and Using a Shareable Parallel Corpus for Machine Translation Evaluation. *IN: Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC-04). Workshop on the Amazing Utility of Parallel and Comparable Corpora, Genoa, Italy. No pagination.

**Estrella, Paula.** (2008). *Evaluating Machine Translation in Context: Metrics and Tools*. PhD Thesis. University of Geneva. [Online]. Available from: < http://www.mt-archive.info/MTS-2007-Estrella-1.pdf>. [Accessed 10 July 2009].

**Estrella, Paula, Olivier Hamon and Andrei Popescu-Belis.** (2007). How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics. *IN: Proceedings of Machine Translation Summit XI*. Copenhagen, Denmark. pp. 167-174.

**European Commission**. (2005). Key Data on Education in Europe [Online]. Available from: <http://eacea.ec.europa.eu/ressources/eurydice/pdf/052EN/007_chapC_052EN.pdf> [Accessed 16 February 2009].

**Falkedal, Kirsten.** (1994). *Evaluation methods for machine translation systems: An historical overview and critical account*, ISSCO draft report, University of Geneva, Geneva.

**FEMTI - A Framework for the Evaluation of Machine Translation in ISLE**. (Homepage). [Online]. Available from: <http://www.issco.unige.ch:8080/cocoon/femti/st-home.html> [Accessed 10 March 2009].

**Fiederer, Rebecca and Sharon O'Brien**. (2009). Quality and Machine Translation: A realistic objective? *Journal of Specialised Translation*, 11, pp. 52-74. [Online]. Available from: < http://www.jostrans.org/issue11/art_fiederer_obrien.pdf> [Accessed 15 May 2009].

**Flanagan, Mary.** (1994). Error classification for MT evaluation. *IN: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas, Maryland, USA*. (AMTA-94). pp. 65-72.

**Flanagan, Mary.** (1997). MT Today: Emerging Roles, New Successes. *Machine Translation,* 12 (1-2), pp. 25-27.

**Flanagan, Marian and Dorothy Kenny.** (2007). Investigating Repetition and Reusability of Translations in Subtitle Corpora for use with Example-Based Machine Translation. *IN: Proceedings of Corpus Linguistics*. Birmingham, UK. [Online]. Available from: <http://www.corpus.bham.ac.uk/corplingproceedings07/paper/129_Paper.pdf> [Accessed 10 May 2009].

**Frey, Lawrence R., Carl H. Botan and Gary L. Kreps.** (1991). *Investigating Communication – An Introduction to Research Methods*. Englewood Cliffs, New Jersey: Prentice Hall.

**Fries-Gedin, Lena.** (2002). Dunkare, Klonken och den gyllene Kvicken: translating the *Harry Potter* phenomenon into Swedish Translation. [Online]. Available from**:** **<**http://www.swedishbookreview.com/old/2002s-gedin.html**>** [Accessed 19 February 2009].

**Gale, William A. and Kenneth W. Church.** (1991). A Program for Aligning Sentences in Bilingual Corpora. *IN: Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, (ACL-91). Berkeley, CA. pp. 177-184.

**Gambier, Yves.** (2003). Screen Transadaptation: Perception and Reception. *The Translator*, 9 (2), pp. 171-189.

**Gambier, Yves.** (ed.) (2004). Traduction audiovisuelle. Audiovisual translation. *Meta*, (special issue), 49 (1), pp. 1-11.

**Gambier, Yves.** (2006). Le sous-titrage: une traduction sélective? *IN:* Jurma Tommola and Yves Gambier (eds.) *Translation and Interpreting. Training and Research*. Turku: University of Turku. pp. 21-37.

**Gambier, Yves.** (2008). Recent developments and challenges in audiovisual translation research *IN:* D. Chiaro, C. Heiss and C. Bucaria (eds.) *Between Text and Image*. Amsterdam/Philadelphia: John Benjamins Publishing. pp 11-33.

**Gamma, Erich, Richard Helm, Ralph Johnson and John M. Vlissides.** (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

**Gebremedhin, Tesfa and Luther Tweeten.** (1994). *Research Methods and Communication in the Social Sciences*. Westport, USA: Praeger.

**Genette, Gérard.** (1997). *Palimpsets: Literature in the second degree.* Channa Newman and Claude Doubinsky (trans.) Lincoln NE/London: University of Nebraska Press.[First published in French as *Palimpsestes* (1982)].

**Global Engineering Education Exchange.** (2009). (Homepage). [Online]. Available from: <http://www.iie.org/programs/global-e3/> [Accessed 6 February 2009].

**Global Translation Systems**. (Homepage). [Online]. Available from: <http://www.globaltranslation.com/> [Accessed 10 March 2009].

**Goldschmidt, Evelyn P.** (1943). *Mediaeval Texts and Their First Appearance in Print*. Oxford: Oxford University Press.

**Göritz, Anja.** (2006). Incentives in Web Studies Methodological Issues and a Review. *International Journal of Internet Science*, 1 (1), pp. 58-70.

**Gottlieb, Henrik.** (1992). Subtitling – a new university discipline. *IN:* C. Dollerup and A. Loddegaard (eds.) *Teaching Translation and Interpreting.* Amsterdam/Philadelphia: John Benjamins Publishing. pp. 161-170.

**Gottlieb, Henrik.** (1994). Subtitling: Diagonal translation. *Perspectives* 2 (1), pp. 101-121.

**Gottlieb, Henrik**. (1995). Establishing a Framework for a Typology of Subtitle Reading Strategies: Viewer Reactions to Deviations from Subtitling Standards. *Translatio*. pp.388-409.

**Gottlieb, Henrik.** (1997). *Subtitles, Translation & Idioms*. PhD Thesis. University of Copenhagen: Centre for Translation Studies and Lexicography.

**Gottlieb, Henrik.** (2005). Multidimensional Translation: Semantics Turned Semiotics. *IN: Proceedings of Multidimensional Translation (MuTra) – EU High Level Scientific Conference Series*, Saarbrücken, Germany. pp. 1-29.*
(Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.) [Online]. Available from: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_Gottlieb_Henrik.pdf> [Accessed 10 March 2009].

**Gough, Nano and Andy Way.** (2004a). Robust Large-Scale EBMT with Marker-Based Segmentation. *IN: Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, Baltimore, MD. pp. 95-104.

**Gough, Nano and Andy Way.** (2004b). Example-Based Controlled Translation. *IN: Proceedings of the 9th Workshop of the European Association for Machine Translation (EAMT-04)*, Valetta, Malta. pp. 73-81.

**Gough, Nano.** (2005). *Example-based machine translation using the marker hypothesis*. PhD Thesis. Dublin City University.

**Green, Thomas.** (1979). The necessity of syntax markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18 (4), pp. 481-496.

**Groves, Declan and Andy Way.** (2005). Hybrid example-based SMT: the best of both worlds? *IN: ACL-2005*: *Workshop on Building and Using Parallel Texts – Data-driven machine translation and beyond*. University of Michigan, Ann Arbor. pp. 183-190.

**Groves, Declan and Andy Way.** (2006a). Hybrid Data-Driven Models of Machine Translation. *Machine Translation, Special Issue on EBMT*, 19 (3-4), pp. 301-323.

**Groves, Declan and Andy Way.** (2006b). Hybridity in MT: Experiments on the Europarl Corpus. *IN: Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo, Norway. pp. 115–124.

**Groves, Declan.** (2007). *Hybrid Data-Driven Models of Machine Translation*. PhD Thesis. Dublin City University.

**Groves, Declan.** (2008). Bringing humans into the loop: Localisation with Machine Translation at Traslán. *IN: Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-08),* Waikiki, HI, USA. pp. 1-12.* (This was a keynote speech at AMTA-2008, and therefore there are no page numbers provided. The page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.)

**Halliday, T.C.** As cited in Van Slype (no further details available).

**Halliday, T.C. and E.A. Briss.** (1977). *The Evaluation and Systems Analysis of the Systran Machine Translation System.* Report RADC-TR-76-399, Rome Air Development Center, Griffiss Air Force Base, NY.

**Hardmeier, Christian and Martin Volk.** (2009). *Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles. IN: Proceedings of Nodalida.* Odense, Denmark. pp. 1-8.* (Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication).

**Hassan, Hany, Yanjun Ma and Andy Way.** (2007). MaTrEx: the DCU Machine Translation System for IWSLT 2007. *IN: Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy. pp. 69-75.

**Hofstadter, Douglas.** (1997). *Le Ton beau de Marot: In Praise of the Music of Language.* New York: Basic.

**Hovy, Eduard, Margaret King and Andrei Popescu-Belis.** (2002a). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17 (1), pp. 43-75.

**Hovy, Eduard, Margaret King and Andrei Popescu-Belis**. (2002b) An Introduction to MT Evaluation. *IN: Handbook of the LREC 2002 Workshop Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*, Las Palmas de Gran Canaria, Spain. pp.1-7.

**Hovy, Eduard.** (1988). *Generating Natural Language under Pragmatic Constraints*. Hillsdale, NJ: Lawrence Erlbaum Associates.

**Huang, Edgar.** (2007). A DVD Dilemma: Ripping for Teaching. *Convergence: The International Journal of Research into New Media Technologies,* 13 (2), pp.129-141.

**Huijsen, Willem-Olaf.** (1998). Controlled language – An Introduction. *IN: Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA. pp. 1-15.

**Hutchins, John.** (1986). *Machine Translation: Past, Present, Future.* West Sussex, England: Ellis Horwood Ltd.

**Hutchins, John.** (1987). Prospects in machine translation. *IN: MT Summit manuscripts and program*, Hakone Prince Hotel, Japan. pp. 48-52.

**Hutchins, John.** (1993). JEIDA report on evaluation methodology. *MT News International: Newsletter of the International Association for Machine Translation*. Issue 4, pp. 24-26.

**Hutchins, John.** (1997). Translation Technology and the Translator. *IN: Proceedings of the Eleventh Conference of the Institute of Translation and Interpreting*. London: ITI. pp. 113-120.

**Hutchins, John.** (1998). The origins of the translator's workstation. *Machine Translation*, 13 (4), pp. 287-307.

**Hutchins, John.** (1999). The development and use of machine translation systems and computer-based translation tools. *IN: Proceedings of the International Conference on Machine Translation & Computer Language Information Processing*. Beijing: Research Center of Computer & Language Engineering, Chinese Academy of Sciences. pp. 1-16.

**Hutchins, John.** (2003a). The (in)famous ALPAC report. *IN*: S. Nirenburg, H. Somers and Y. Wilks (eds.) *Readings in Machine Translation*. Cambridge, Massachusetts/London, England: MIT Press. pp. 131-136.

**Hutchins, John.** (2003b). Commercial Systems: the state of the art *IN:* H. Somers (ed.) *Computers and Translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins Publishing. pp. 161-174.

**Hutchins, John.** (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, 17 (1-2), pp.5-38.

**Hutchins, John.** (2006). Example-based Machine Translation – a review and commentary. *Machine Translation*, 19 (3-4), pp. 197-211.

**Hutchins, John and Harold Somers.** (1992). *An Introduction to Machine Translation.* London: Academic Press.

**Huysmans, Frank and Jos de Haan.** (2001). Media en ICT: Omgaan met een Overvloedig Aanbod *IN:* K. Breedveld and A. van den Broek (eds.) *Trends in de Tijd: Een Schets van Recente Ontwikkelingen in Tijdsbesteding en Tijdsordening*. The Hague: Sociaal en Cultureel Planbureau. pp. 75–95

**ISLE.** (1999). *Information Society Technologies (IST): Final Report*. IST-1999-10647. [Online]. Available from: <http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm> [Accessed 10 March 2009].

**ISO/IEC,** (1999). *ISO/IEC 14598-1: 1999. Information technology – software product evaluation – Part 1: General overview*, Geneva: International Organization for Standardization & International Electrotechnical Commission.

**Ivarsson, Jan and Mary Carroll.** (1998). *Subtitling*. Simrishamn: TransEdit.

**Ivarsson, Jan.** (1992). *Subtitling for the media: A handbook of an art*. Stockholm: Transedit.

**Jakobson, Roman.** (1959) On Linguistic aspects of Translation *IN:* Reuben Brower (ed.) *On Translation.* Cambridge, Massachusetts: Harvard University Press. (Reprinted in Venuti (ed.) 2000, pp. 113-118.)

**James, Heulwen.** (2001). Quality Control of Subtitles: Review or Preview? *IN:* Y. Gambier and H. Gottlieb (eds.) *(Multi) Media Translation.* Amsterdam/Philadelphia: John Benjamins Publishing. pp. 151-160.

**JEIDA.** (1992). *JEIDA Methodology and Criteria on Machine Translation Evaluation* (JEIDA Report). H. Nomura (ed.) Japan Electronic Industry Development Association.

**Kay, Martin.** (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12 (1-2), pp. 3-23.

**Kennedy, Graham.** (1998). *An Introduction to Corpus Linguistics.* London/New York: Longman.

**Kenny, Dorothy.** (2001). *Lexis and Creativity in Translation: a corpus-based study*. Manchester, England/Northampton, M.A., USA: St. Jerome.

**King, Margaret.** (1996). The Evaluation of Natural Language Processing Systems. *IN: Special edition of Communications of the ACM on Natural Language Processing*, 39 (1), pp 73-79.

**King, Margaret.** (1997). Evaluating Translation. *IN*: C. Hauenschild and S. Heizmann (eds.). *Machine Translation: Translation Theory*. Berlin/New York: Mouton de Gruyter. pp. 251-263.

**King, Margaret, Eduard Hovy, John White, Benjamin K. T'sou and Yusoff Zaharin.** (1999). MT Evaluation. *IN: Proceedings of the Machine Translation Summit VII*, Singapore. pp. 197-207.

**King, Margaret, Andrei Popescu-Belis and Eduard Hovy.** (2003). FEMTI: creating and using a framework for the MT evaluation. *IN: Proceedings of the Machine Translation Summit IX,* New Orleans, LA, USA. pp. 224-231.

**Klare, George R.** (1977). Readable technical writing: some observations. *Technical Communication*, 2nd Quarter, 24, pp. 1-3. (Reprinted in C. Harkins and D.L. Plung (eds.) (1982). *A guide for writing better technical papers.* New York: IEEE Press. pp. 149-151).

**Knight, Kevin.** (1997). Automating Knowledge Acquisition for Machine Translation. *AI Magazine*, 18 (4), pp. 81-96.

**Koehn, Philip.** (2004). Pharaoh: A beam search decoder for phrase-based Statistical Machine Translation models. *IN: Proceedings of the Workshop on Machine Translation: From real users to research, (AMTA-04).* pp. 115-124.

**Koehn, Phillip.** (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *IN: Proceedings of the Machine Translation Summit X*, Phuket, Thailand. pp. 79-86.

**Koehn, Phillip, Franz Josef Och and Daniel Marcu.** (2003). Statistical Phrase-based Translation. *IN: Proceedings of the Human Language Technology conference.* ACL: Edmonton, Canada. pp. 48-54.
**Koehn, Philip and Christof Monz.** (2006). Manual and automatic evaluation of machine translation between European languages. *IN: Proceedings of the NAACL 2006 Workshop on Statistical Machine Translation.* pp. 102-121.

**Koolstra, Cees M., Allerd L. Peeters and Herman Spinhof.** (2002). The pros and cons of dubbing and subtitling. *European Journal of Communication*, (17), 3, pp. 325-354.

**Kress, Gunther and Theo van Leeuwen.** (2001). Multimodal Discourse - *The Modes and Media of Contemporary Communication.* London: Arnold.

**Kristeva, Julia.** (1980a). The Bounded Text *IN:* Léon Roudiez (ed.) *Desire in Language: A Semiotic Approach to Literature and Art.* New York: Columbia University Press, London: Basil Blackwell. pp. 36-63.

**Kristeva, Julia.** (1980b). Word, Dialogue, Novel *IN:* Léon Roudiez (ed.) *Desire in Language: A Semiotic Approach to Literature and Art.* New York: Columbia University Press, London: Basil Blackwell. pp. 64-91.

**Kulesza, Alex and Stuart M. Shieber.** (2004). A Learning Approach to Improving Sentence-Level MT Evaluation. *IN: Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI).* Baltimore, MD, USA. pp. 75-84.

**Lapata, Mirella**. (2006). Automatic Evaluation of Information Ordering: Kendall's Tau. *Computational Linguistics*, 3 (4), pp. 471-484.

**Landis, R. J. and Koch, G. K.** (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159-174.

**Lavecchia, Caroline, Kamel Smaili and David Langlois.** (2007). Building parallel corpora from movies. *IN: Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science,* Funchal, Madeira. No pagination.

**Lavie, Alon and Abhaya Agarwal.** (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgements. *IN: Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07),* Prague, Czech Republic. pp. 228-231.

**Lavie, Alon, Kenji Sagae and Shyamsundar Jayaraman.** (2004). The significance of recall in automatic metrics for MT evaluation. *IN: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-04),* Washington, DC, USA. pp. 134-143.

**LDC.** (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

**Leusch, Gregor, Nicola Ueffing and Hermann Ney.** (2006). CDER: Efficient MT evaluation using block movements. *IN: Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL).* pp. 241-248.

**Likert, Rensis.** (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140, pp. 1-55.

**Liu, Ding and Daniel Gildea.** (2005). Syntactic features for evaluation of machine translation. *IN: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-05),* Ann Arbor, Michigan, USA. pp. 25-32.

**Loffler-Laurian, Anne-Marie.** (1983). Pour une typologie des erreurs dans la traduction automatique. *Multilingua,* 2 (2), pp. 65-78.

**Loffler-Laurian, Anne-Marie.** (1996). *La Traduction Automatique*. Villeneuve d'Ascq. Presses Universitaires du Septentrion.

**López Ciruelos, A.** (2003). Una defensa crítica de las memorias de traducción. *Panace*, 4 (12), pp. 180–82.

**Luyken, Geory-Michael, Thomas Herbst, Jo Langham-Brown, Helene Reid and Herman Spinhof.** (1991). *Overcoming language barriers in television: dubbing and subtitling for the European audience*. Manchester: European Institute for the Media.

**Ma, Yanjun, John Tinsley, Hany Hassan, Jinhua Du and Andy Way.** (2008). Exploiting Alignment Techniques in MaTrEx: the DCU MT System for IWSLT 2008. *IN: Proceedings of the IWSLT 2008 Workshop*, Honolulu, HI. pp. 26-33.

**Marleau, Lucien.** (1982). Les sous-titres…un mal nécessaire. *Meta* 27 (3), pp. 271-285.

**Mayoral, Asensio Roberto, Dorothy Kelly and Natividad Gallardo.** (1988). Concept of constrained translation. Non-linguistic perspective of translation. *Meta* 33 (3), pp. 356-367.

**McEnery, Tony and Andrew Wilson.** (2001). *Corpus Linguistics.* 2nd edition. Edinburgh: Edinburgh University Press.

**Melamed, I. Dan, Ryan Green and Joseph P. Turian.** (2003). Precision and recall of Machine Translation. *IN: Proceedings of Human Language Technology North American Chapter of the Association of Computational Linguistics Conference*. pp. 61-63.

**Melby, Alan.** (1997). Some notes on the proper place of men and machines in language translation. *Machine Translation*, 12 (1-2), pp. 29-34.

**Melero, Maite, Antoni Oliver and Toni Badia.** (2006). Automatic multilingual subtitling in the eTITLE project. *IN: Proceedings of Translating and the Computer 28,* London, England: ASLIB. No pagination.

**Mima, Hideki, Hitoshi Iida and Osamu Furuse.** (1998). Simultaneous Interpretation Utilizing Example-based Incremental Transfer. *IN: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-98*). pp. 855-861.

**Moore, Nick.** (2006). *How to do research: a practical guide to designing and managing research projects*. 3rd edition. Facet: London

**Morrissey, Sara and Andy Way.** (2005). An example-based approach to translating sign language, *IN: Proceedings of Second Workshop on Example-Based Machine Translation*, MT Summit X, Phuket, Thailand, pp. 109-116.

**Morrissey, Sara and Andy Way.** (2006). Lost in translation: the problems of using mainstream MT evaluation metrics for sign language translation. *IN: Proceedings to the 5th International Conference on Language Resources and Evaluation* (5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages"), Genoa, Italy. pp. 91-98.

**Mossop, Brian.** (2006). Has Computerization Changed Translation? *Meta* 51 (4), pp. 787-793.

**Nagao, Makoto.** (1989). *Machine Translation: How far can it go?* Oxford, UK: Oxford University Press. Translated by Norman Cook.

**Nagao, Makoto.** (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *IN: Proceedings of the international NATO symposium on Artificial and human intelligence*, October 1981. Lyon, France: Elsevier North-Holland, Inc. pp. 173-180.

**National Science Foundation.** (2008). (Homepage). [Online]. Available from: <http://www.nsf.gov/statistics/seind08/c2/c2h.htm#c2sh3> [Accessed 6 February 2009].

**Nedoma, Andrew and Jurek Nedoma.** (2004). Problems with CAT tools related to translations into Central and Eastern European languages. *IN: Proceedings of Translating and the Computer 26*. London. No pagination.

**NHK Annual Report.** (1994). [Online]. Available from: <http://www.nhk.or.jp/strl/results/annual94/index.html> [Accessed 10 May 2009].

**NHK Annual Report.** (1995). [Online]. Available from: <http://www.nhk.or.jp/strl/results/annual95/index.html> [Accessed 10 May 2009].

**NHK Annual Report.** (2002). [Online]. Available from: < http://www.nhk.or.jp/strl/results/annual2002/index.html> [Accessed 10 May 2009].

**NHK Annual Report.** (2003). [Online]. Available from: < http://www.nhk.or.jp/strl/results/annual2003/index.html> [Accessed 10 May 2009].

**NHK Annual Report.** (2004). [Online]. Available from: < http://www.nhk.or.jp/strl/results/annual2004/index.html> [Accessed 10 May 2009].

**Nirenburg, Sergei, Harold Somers and Yorick Wilks.** (eds.) (2003). *Readings in Machine Translation*. Cambridge, MA: MIT Press.

**Nomura, Hirosato and S. Ishara.** (1992). Methodology and criteria on machine translation evaluation. *IN: MT Evaluation: Basis for future directions. Proceedings of a workshop sponsored by the National Science Foundation*, San Diego, California. pp. 11-12.

**O'Brien, Sharon.** (2006). *Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD Thesis. Dublin City University.

**O'Hagan, Minako.** (2003a). Middle Earth Poses Challenges to Japanese Subtitling. *The Localization Industry Standards Association Newsletter XII*, 1.5. [Online]. Available from: <http://www.lisa.org/globalizationinsider/2003/03/middle_earth_po.html> [Accessed 10 May 2009].

**O'Hagan, Minako.** (2003b). Can Language Technology Respond to the Subtitler's Dilemma? *IN: Proceedings of the 25th International Conference on Translating and the Computer*, November. London: ASLIB. pp. 1-18. *(Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.)

**O'Hagan, Minako.** (2007). Impact of DVD on Translation: Language options as an essential add-on feature. *Convergence*, 13 (2), pp. 157-168.

**O'Halloran, Kay.** (2004). *Multimodal Discourse Analysis: Systemic Functional Perspectives*. London/New York: Continuum International Publishing Group.

**Och, Franz Josef, Christoph Tillmann and Hermann Ney.** (1999). Improved alignment models for statistical machine translation. *IN: Proceedings of the Joint Conference of Empirical Methods in Natural language Processing and Very Large Corpora*. University of Maryland, MD, USA. pp. 20-28.

**Och, Franz Josef and Hermann Ney.** (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29 (1), pp. 19-51.

**Olohan, Maeve.** (2004). *Introducing Corpora in Translation Studies*. London/New York: Routledge.

**Oppenheim, Abraham Naftali.** (1966). *Questionnaire Design and Attitude Measurement*. London: Heinemann.

**Orero, Pilar.** (2004). Audiovisual translation: A new dynamic umbrella. *IN: Topics in Audiovisual Translation*. Orero, P. (ed.) Amsterdam/Philadelphia: John Benjamins Publishing. pp. vii-xiii.

**Orr, D and V. Small.** (1967). Comprehensibility of Machine-aided Translations of Russian Scientific Documents *Mechanical Translation and Computational Linguistics* (10), pp. 1-10.

**Ortiz-Martínez, Daniel, Ismael García-Varea and Francisco Casacuberta.** (2005). Thot: A toolkit to train phrase-based statistical translation models. *IN: Proceedings of Machine Translation Summit X*, Phuket, Thailand. pp. 141-148.

**Owczarzak, Karolina, Declan Groves, Josef van Genabith and Andy Way.** (2006). Contextual bitext-derived paraphrases in automatic MT evaluation. *IN: Proceedings of the HLT-NAACL 2006 Workshop on Statistical Machine Translation*. pp. 86-93.

**Owczarzak, Karolina.** (2008). *A Novel Dependency-based Evaluation Metric for Machine Translation*. PhD Thesis. Dublin City University.

**Ozdowska, Sylwia and Andy Way**. (2009). Optimal Bilingual Data for French-English PB-SMT. *IN: Proceedings of the 13th Annual Meeting of the European Association for Machine Translation(EAMT-09)*. Barcelona, Spain. pp. 96-103.

**Pallant, Julie.** (2005). *SPSS Survival Manual*. 2nd edition. Open University Press: NY.

**Pankowicz, Z.L.** (1978). Evaluation of machine translation: a position paper - Luxembourg, CEC, memorandum. pp. 6.

**Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu.** (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *IN: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02),* Philadelphia, USA. pp. 311-318.

**Paul, Michael.** (2006). Overview of the IWSLT 2006 evaluation campaign. *IN: Proceedings of the International Workshop on Spoken Language Translation*, pp. 1-15.

**Pierce, John R., John B. Carroll, Eric P. Hamp, David G. Hays, Charles F. Hockett, Anthony G. Oettinger and Alan Perlis.** (1966). Language and machines computers in translation and linguistics. *Technical report, Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, National Research Council*, Washington, DC, USA.

**Piperidis, Stelios, Iason Demiros, Prokopis Prokopidis.** (2004). Multimodal Multilingual Information Processing for Automatic Subtitle Generation: Resources, Methods and System Architecture (MUSA). *Presentation at E-Tools and Translation session*, Languages and the Media, Berlin, Germany. [Online]. Available from: <http://sifnos.ilsp.gr/musa/LM> [Accessed 10 March 2009].

**Piperidis, Stelios, Iason Demiros, Prokopis Prokopidis.** (2005). Infrastructure for a multilingual subtitle generation system *IN: 9th International Symposium on Social Communication*, Santiago de Cuba, Cuba. no pagination.

**Popowich, Fred, Paul McFetridge, Davide Turcato and Janine Toole.** (2000). Machine Translation of Closed Captions. *Machine Translation*, 15 (4), pp. 311-341.

**Quah, Chiew Kin.** (2006). *Translation and Technology*. Basingstoke and New York: Palgrave MacMillan.

**Rapley, Tim.** (2004). Interviews. *IN:* C. Seale, G. Gobo, J.F. Dubrium and D. Silverman (eds.) *Qualitative Research Practice*. London/Thousand Oaks/New Delhi: Sage Publications. pp. 15-33.

**Reid, Helene.** (1978). Subtitling, the intelligent solution. *IN:* P.A. Horguelin (ed.), *La traduction, une profession–Translating, a profession*. Council of Translators and Interpreters of Canada, Ottawa. pp. 421-428.

**Reinke, Uwe.** (2004). *Translation Memories. Systeme – Konzepte – Linguistische Optimierung*. Saarbrücker Beiträge zur Sprach- und Translationswissenschaft. Europäischer Verlag der Wissenschaften. Frankfurt am Main: Peter Lang.

**Resnik, Philip.** (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, pp. 95–130.

**Resnik, Philip and Mona Diab.** (2000). Measuring verb similarity. *IN: Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 399–404.

**Rosenthal, Robert.** (1966). *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

**Roturier, Johann.** (2006). *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users*. PhD Thesis. Dublin City University.

**Sadler, Victor and Ronald Vendelmans.** (1990). Pilot Implementation of a Bilingual Knowledge Bank. *IN: COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland: COLING, Vol. 3, pp. 449-451.

**Sadler, Victor.** (1991). The Textual Knowledge Bank: Design, Construction, Applications *IN: International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP)*, Kyoto, Japan. pp. 17-32.

**Shuttleworth, Mark.** (2002). Combining MT and TM on a Technology-oriented Translation Masters: Aims and Perspectives. *IN: Proceedings of 6th EAMT Workshop on Teaching Machine Translation*, Manchester, UK. pp 123-129.

**Sinaiko, H.W.** (1978). Some thoughts about evaluating language translation. Luxembourg, CEC, memorandum, pp. 7.

**Snover, Matthew, Bonnie Dorr, Richard Schwartz, John Makhoul and Linnea Micciulla.** (2006). A study of translation error rate with targeted human annotation. *IN: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06),* Cambridge, Massachusetts, USA. pp. 223-231.

**Somers, Harold and Elizabeth Wild.** (2000). Evaluating Machine Translation: the Cloze procedure revisited. *IN: Proceedings of Translating and the Computer 22*. London, UK. No pagination.

**Somers, Harold.** (1999). Review article: Example-based Machine Translation. *Machine Translation*, 14, pp. 113-157.

**Somers, Harold.** (2003). An Overview of EBMT *IN:* M. Carl and A. Way (eds.) *Recent Advances in Example-Based Machine Translation*. Dordrecht/Boston/London: Kluwer Academic Publishers. pp. 3-57.

**Srivastava Ankit Kumar, Rejwanul Haque, Sudip Kumar Naskar and Andy Way.** *(*2008). MaTrEx: the DCU MT System for ICON 2008. *IN: Proceedings of the NLP Tools Contest: Statistical Machine Translation (English to Hindi), 6th International Conference on Natural Language Processing*, Pune, India (in press).

**Stroppa, Nicolas and Andy Way.** (2006). MaTrEx: DCU Machine Translation System for IWSLT 2006. *IN: Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan. pp. 31-36.

**Stroppa, Nicolas, Declan Groves, Kepa Sarasola and Andy Way.** (2006). Example-based Machine Translation of the Basque Language. *IN: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, Boston, MA. pp. 232-241.

**Sumita, Eiichiro, Hitoshi Iida, and H. Kohyama**. (1990). Translating With Examples: A New Approach to Machine Translation. *IN: Third International Conference on Theoretical and Methodological Issues in Machine Translation*. Austin, TX: Springer Netherlands. pp. 203–212.

**Sumiyoshi, Hideki, Hideki Tanaka, Nobuko Hatada and Terumasa Ehara.** (1995). Translation workbench for generating subtitles for English TV news. n°435, pp. 1-8. NHK Science and Technical Research Laboratories, Tokyo, Japan.

**Swales, John M.** (1990). *Genre Analysis.* Cambridge: Cambridge University Press.

**Taylor, Christopher.** (2006). "I knew he'd say that!" A consideration of the predictability of language use in film. *IN: Proceedings of Multidimensional Translation (MuTra) – EU High Level Scientific Conference Series*, Copenhagen, Denmark. pp. 1-11.* (Pages in this publication are unnumbered; the page numbers cited above have been provided for the sake of convenience and relate only to the pages of this particular article rather than the whole publication.) [Online]. Available from:

<http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Taylor_Christopher.pdf> [Accessed 10 March 2009].

**TEMAA.** (1996). TEMAA Final Report, LRE-62-070, Center for Sprogteknologi, Copenhagen, Denmark. [Online]. Available at: <http://cst.dk/temaa/D16/d16exp.html>. [Accessed 10 March 2009].

**Tiedemann, Jörg.** (2007a). Improved sentence alignment for movie subtitles. *IN: Proceedings of Recent Advances in Natural Language Processing*. Borovets, Bulgaria pp. 582–588.

**Tiedemann, Jörg.** (2007b). Building a multilingual parallel subtitle corpus. *IN: Proceedings of 17th Computational Linguistics in the Netherlands*, Leuven, Belgium. No pagination.

**Tiedemann, Jörg.** (2008). Synchronizing Translated Movie Subtitles. *IN: Proceedings of Language Resources and Evaluation Conference (LREC-08).* Marrakesh, Morocco. No pagination.

**Tinsley, John, Yanjun Ma, Sylwia Ozdowska and Andy Way.** (2008). MaTrEx: the DCU MT System for WMT 2008. *IN: Proceedings of the Third Workshop on Statistical Machine Translation, (ACL 2008)*. Columbus, OH. pp. 171-174.

**Titford, Christopher.** (1982). Subtitling: constrained translation. *Lebende Sprachen* 27 (3), pp. 113-116.

**Trujillo, Arturo.** (1999). *Translation engines: Techniques for Machine Translation*. London: Springer.

**Turcato, Davide and Fred Popowich.** (2003). What is Example-Based Machine Translation? *IN:* M. Carl and A. Way (eds.) *Recent Advances in Example-Based Machine Translation*. Dordrecht/Boston/London: Kluwer Academic Publishers. pp. 59-82. [Revised version of Workshop paper 2001].

**Turian, Joseph, Luke Shen and Dan Melamed.** (2003). Evaluation of machine translation and its evaluation. *IN: Proceedings of the Machine Translation Summit IX*, New Orleans, USA. pp. 386-393.

**UNESCO Institute for Statistics.** (2009). [Online]. Available from: <http://stats.uis.unesco.org/unesco/ReportFolders/ReportFolders.aspx> [Accessed 16 February 2009].

**Van Slype, Georges.** (1979). Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical Report BR19142, Bureau Marcel van Dijk/European Commission (DG XIII), Brussels. [Online]. Available from: <http://www.issco.unige.ch/projects/isle/van-slype.pdf> [Accessed 10 March 2009].

**Vasconcellos, Muriel, Bernard Scott, John Chandiouz and Hideki Tanaka.** (1991). The MT User Experience: Panel Discussion. *IN: Proceedings of MT Summit III*, Washington DC, USA. pp. 121-126.

**Vasconcellos, Muriel.** (1992). Panel: Apples, Oranges, or Kiwis? Criteria for the comparison of MT systems. *IN: MT Evaluation: Basis for future directions. Proceedings of a workshop sponsored by the National Science Foundation*. San Diego, California. pp. 37-50.

**Vauquois, Bernard.** (1968). Structures profondes et traduction automatique. Le système du CETA. *Revue Roumaine de linguistique*, 13, pp. 105-130.

**Ventola, Eija and Anna Mauranen.** (1996). *Academic Writing: Intercultural and Textual Issues*. Amsterdam/Philadelphia: John Benjamins Publishing.

**Venuti, Lawrence.** (ed.) (2000). *The Translation Studies Reader.* London and New York: Routledge.

**Volk, Martin and Søren Harder.** (2007). Evaluating MT with Translations or Translators. What is the Difference? *IN:* Proceedings *of MT Summit XI*, Copenhagen, Denmark. pp. 499-506.

**Volk, Martin.** (2008). The Automatic Translation of Film Subtitles. A Machine Translation Success Story? *IN:* J. Nivre, M. Dahllöf, and B. Megyesi (eds.) *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*. Sweden: Uppsala University. pp. 202-214.

**Wagner, Simone.** (1998). Small-scale evaluation methods. *IN:* R. Nübel and U. Seewald-Heeg (eds.) *Evaluation of the Linguistic Performance of Machine Translation Systems. Proceedings of the Workshop at KONVENS-98,* Bonn, Germany. pp. 93-105.

**Wallis, Julian.** (2006). *Interactive Translation vs. Pre-translation in the Context of Translation Memory Systems: Investigating the effects of translation method on productivity, quality and translator satisfaction*. MA Thesis. University of Ottawa. Translation.

**Way, Andy and Nano Gough.** (2003). wEBMT: Developing and Validating an Example- Based Machine Translation System using the World Wide Web. *Computational Linguistics: Special Issue on the Web as Corpus*, 29 (3), pp. 421–457.

**Way, Andy and Nano Gough.** (2005a). Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11 (3), pp. 295–309.

**Way, Andy and Nano Gough.** (2005b). Controlled Translation in an Example-Based Environment: what do Automatic Evaluation Metrics tell us? *Machine Translation*, 19 (1), pp. 1–36.

**Way, Andy, Nano Gough and Declan Groves.** (2005). Comparing Example-Based and Statistical Machine Translation. [Online]. Slides available from: <http://www.computing.dcu.ie/~away/PUBS/2005/Edinburgh.ppt> [Accessed 10 March 2009].

**Webb, Lynne E.** (1998). *Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis*. MA Thesis. Monterey, CA: Monterey Institute of International Studies. [Online]. Available from: <http://www.tradulex.org/Bibliography/Webb.htm> [Accessed 10 March 2009].
**Weiss, Robert S.** (1994). *Learning from strangers: The art and method of qualitative interviewing*. New York: Free Press.

**Weiss, Sholom M.and Casimir A. Kulikowski**. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann Publishers.

**White, John S.** (2000a). Contemplating automatic MT evaluation. *IN*: J. S. White (ed.). *Envisioning Machine Translation in the Information Future.* Berlin/Heidelberg/New York/Barcelona/Hong Kong/London/Milan/Paris/Singapore/Tokyo: Springer. pp. 100-108.

**White, John S.** (2000b). Toward an Automated, Task-Based MT Evaluation Strategy. *IN*: B. Maegaard (ed.). *Proceedings of the Workshop on Machine Translation Evaluation, LREC*, Athens, Greece. No pagination.

**White, John S.** (2003). How to evaluate machine translation. *IN*: H. Somers (ed.). *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins Publishing. pp. 211-244.

**White, John S. and Kathryn B. Taylor.** (1998). A Task-Oriented Evaluation Metric for Machine Translation. *IN: Proceedings of Language Resources and Evaluation Conference*, Granada, Spain. pp. 21-27.

**White, John S., Theresa O'Connell and Francis O'Mara.** (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *IN*: *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-94).* Washington, DC, USA. pp. 193-205.

**Whitelock, Pete.** (1994). Shake and Bake Translation *IN:* C. J. Rupp, M. A. Rosner and R. L. Johnson (eds.) *Constraints, Language and Computation*. London: Academic Press. pp. 339-359.

**Whyman, Edward and Harold Somers.** (1999). Evaluation metrics for a Translation Memory system. *Software-Practice and Experience* 29, pp. 1265-1284.

**Wilks, Yorick.** (1978). The value of the monolingual component in MT evaluation and its role in the Battelle report on SYSTRAN. Luxembourg, CEC, memorandum.

**Williams, Jenny and Andrew Chesterman.** (2002). *The Map: A beginner's Guide to doing Research in Translation Studies.* Manchester/Northampton, M.A.: St. Jerome.

**Winiwarter, Werner.** (2007). Machine Translation Using Corpus-Based Acquisition of Transfer Rules. *IN: Proceedings of the Second International Conference on Digital Information Management (IEEE Engineering Management Society)*. pp. 345-350.

**Womeng** (2005) (Homepage). [Online]. Available from: <http://www.womeng.net/intro.htm> [Accessed 6 February 2009].

**Yagoda, Gerald and William Wolfson.** (1964). Examiner influence on projective test responses. *Journal of Clinical Psychology*, 20, pp. 398.

**Yamada, Kenji and Kevin Knight.** (2001). A Syntax-Based Statistical Translation Model. *IN: Proceedings of 39th Annual meeting of the Association for Computational Linguistics (ACL-01).* Toulouse, France: Morgan Kaufmann Publishers. pp. 523–530.

**Zabalbeascoa, Patrick.** (1997). Dubbing and the nonverbal dimension of translation. *IN:* F. Poyatos (ed.) *Nonverbal Communication and Translation*. Amsterdam/Philadelphia: John Benjamins Publishing. pp. 327-342.

# - Filmography

# Filmography

†: These movies were included in the pilot study but not in the current study.

*2001: A Space Odyssey*. (1968). [DVD]. USA: MGM.
*As good as it gets*. (1997). [DVD]. USA: TriStar pictures.
*Being John Malkovich*. (1999). [DVD]. Puerto Rico: Gramercy Pictures.
† *Breakfast at Tiffany's*. (1961). [DVD]. USA: Jurow-Shepherd.
*Casablanca*. (1942). [DVD]. USA: Warner Bros. Pictures.
*Cast Away*. (2000). [DVD]. USA: DreamWorks SKG.
*Chinatown*. (1974). [DVD]. USA: Paramount Pictures.
*Dr. Strangelove*. (1964). [DVD]. USA: Hawk Films.
*Easy Rider*. (1969). [DVD]. USA: Columbia Pictures Corporation.
*The End of the Affair*. (1999). [DVD]. Canada: Columbia Pictures Corporation.
*Frantic*. (1988). [DVD]. USA: Warner Bros. Pictures.
† *Frenzy*. (1972). [DVD]. USA: Universal Pictures.
*Funny Face*. (1957). [DVD]. USA: Paramount Pictures.
*Get Carter*. (2000). [DVD]. USA: Morgan Creek Productions.
*Harry Potter and the Philosopher's Stone*. (2001). [DVD]. UK: 1492 Pictures.
*Harry Potter and the Chamber of Secrets*. (2002). [DVD]. UK: 1492 Pictures.
*Harry Potter and the Prisoner of Azkaban*. (2004). [DVD]. UK: Warner Bros. Pictures.
*Harry Potter and the Goblet of Fire*. (2005). [DVD]. UK: Warner Bros. Pictures.
*Heat*. (1995). [DVD]. USA: Warner Bros. Pictures.
† *Instinct*. (1999). [DVD]. USA: Spyglass Entertainment.
*Kill Bill volume 1*. (2003). [DVD]. USA: Miramax Films.
*Kill Bill volume 2*. (2004). [DVD]. USA: Miramax Films.
*Panic Room*. (2002). [DVD]. USA: Columbia Pictures Corporation.
*Paris when it sizzles*. (1964). [DVD]. USA: Richard Quine Productions.
*Pretty Woman*. (1990). [DVD]. USA: Touchstone Pictures.
*Sabrina*. (1954). [DVD]. USA: Paramount Pictures.
*Sex, Lies and Videotapes*. (1989). [DVD]. USA: Outlaw Productions.
† *Shakespeare in Love*. (1998). [DVD]. USA: Universal Pictures.
† *This is Spinal Tap*. (1984). [DVD]. USA: Spinal Tap Productions.
† *Singing in the Rain*. (1952). [DVD]. USA: MGM.
† *Spartacus*. (1960). [DVD]. USA: Bryna Productions.
*Sunset Boulevard*. (1950). [DVD]. USA: Paramount Pictures.
† *Terminator 2: Judgement Day*. (1991). [DVD]. USA: Carolco Pictures.
*The Green Mile*. (1999). [DVD]. USA: Castle Rock Entertainment.
*The Lord of the Rings: The Fellowship of the Ring*. (1999). [DVD]. New Zealand: New Line Cinema.
*The Lord of the Rings: The Two Towers*. (2002). [DVD]. New Zealand: New Line Cinema.
*The Lord of the Rings: Return of the King*. (2003). [DVD]. New Zealand: New Line Cinema.
*The Matrix Revolutions*. (2003). [DVD]. USA: Warner Bros. Pictures.
*Thelma and Louise*. (1991). [DVD]. USA: MGM.
† *To Kill a Mockingbird*. (1962). [DVD]. USA: Brentwood Productions.
*Vertigo*. (1958). [DVD]. USA: Alfred J. Hitchcock Productions.
† *Willy Wonka and the Chocolate Factory*. (1971). [DVD]. USA: David L. Wolper Productions.
† *Wizard of OZ*. (1939). [DVD]. USA: MGM.

# – Appendices

# APPENDIX A

A breakdown of the results is presented below, based on each of the ten DVD movie clips shown to the subjects to gauge whether or not the German subtitles we intended to use in our corpora were of an acceptable quality. If the subjects added in extra comments during the interview questionnaire, these are reported on following each Table respectively. For each scale 6 is the best score and 1 is the worst. '×' in the errors category means there were no errors detected and '✓' means errors were detected.

**Table A-1**: German subtitles and English soundtrack: As Good as it Gets

| **Clip 1** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | ✓ |
| **Answered internal checks** | ✓✓ | ✓✓ | ✓✓ |
| **Channels used** | Image, Soundtrack | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles |
| **Comprehensibility** | 6 | 6 | 5 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | | × | × |
| **Satisfaction** | 5 | 5 | 5 |

**Comments:** Subjects did not notice any particular errors; however, they all said they would probably translate one or two subtitles differently. They would not fully agree with some translations, which is simply a difference of opinion.

**Table A-2:** German subtitles and Italian soundtrack: Frantic

| **Clip 2** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | Maybe |
| **Answer internal checks** | ✓✓ | ✓✓ | ✓× |
| **Soundtrack Image Subtitles** | Image, Subtitles | Image, Subtitles | Image, Subtitles |
| **Comprehensibility** | 6 | 4-5 | 5 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 5 | 5 | 5 |
| **Recommend subtitles for entire movie on DVD** | ✓ | ✓ | ✓ |

**Comments:** Subjects said subtitles were acceptable to be used on a DVD, but had some reservations about a few linguistic choices including the use of informal and formal language. They all mentioned that they have no knowledge of Italian so they are not in a position to comment on the correct rendering of the original soundtrack in the German subtitles. The third subject made the point that as a viewer she would be very happy with the subtitles, but as a linguist she would not be as satisfied, giving them a lower rating.

**Table A-3:** German subtitles and English soundtrack: The Green Mile

| **Clip 3** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | ✓ |
| **Answer internal checks** | ✓✓ | ✓✓ | ✓✓ |
| **Soundtrack Image Subtitles** | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles |
| **Comprehensibility** | 6 | 5 | 6 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 6 | 4 | 6 |

**Comments:** Subject 2 thought the subtitles contained unusual sentence structure (subordinate clauses), swearing when there was no need for this type of language and incorrect style in places.

**Table A-4:** German subtitles and Italian soundtrack: Get Carter

| **Clip 4** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | ✓ |
| **Answer internal checks** | ✓× | ✓✓ | ✓✓ |
| **Soundtrack Image Subtitles** | Image, Subtitles | Image, Subtitles | Image, Subtitles |
| **Comprehensibility** | 5 | 6 | 6 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 6 | 6 | 6 |
| **Recommend subtitles for entire movie on DVD** | ✓ | ✓ | ✓ |

**Comments:** Subject 1 rated the subtitles 5 for comprehensibility, stating that any lack of comprehension was to do with the clip being out of context, and not necessarily a fault with the subtitles. In this clip the dialogue was quite fast, and therefore the subtitles were also changing quite often to keep up with the characters. Subjects 1 and 2 commented on this point in particular.

**Table A-5:** German subtitles and English soundtrack: Harry Potter and the Prisoner of Azkaban

| Clip 5 | Subject 1 | Subject 2 | Subject 3 |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | Not sure |
| **Answer internal checks** | ×✓ | ×✓ | ×✓ |
| **Soundtrack Image Subtitles** | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles |
| **Comprehensibility** | 6 | 4 | 5-6 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 5 | 5 | 5 |

**Comments:** Overall the subjects were happy with the subtitles provided, although background noises and sound quality were mentioned as two factors that affected how satisfied they were. All subjects thought that the style of the language was exceptionally good, as it is often difficult to translate the language of adolescents and teenagers.

**Table A-6:** German subtitles and Italian soundtrack: Casablanca

| Clip 6 | Subject 1 | Subject 2 | Subject 3 |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | ✓ |
| **Answer internal checks** | ✓✓ | ✓✓ | ✓✓ |
| **Soundtrack Image Subtitles** | Image, Subtitles | Image, Subtitles | Image, Subtitles |
| **Comprehensibility** | 2-3 | 5 | 5-6 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | ✓ | × | × |
| **Satisfaction** | 2 | 5 | 5 |
| **Recommend subtitles for entire movie on DVD** | × | ✓ | ✓ |

**Comments:** There were mixed views between the subjects in relation to the comprehensibility of and satisfaction with these subtitles. Subject 1 thought the subtitles were not very convincing, and understood most of the story using previous knowledge.

This subject found that it was a lot of effort to understand the dialogue. She also found the subtitles were not very idiomatic. In addition to this the subtitles were slightly stilted and artificial. She noted one spelling error, using 'Sie' (formal you) instead of 'sie' (her). However, subject 3 found the use of the subjunctive in the subtitles quite strange and the fact that in the subtitles they use 'Monsieur' instead of 'Mr.' or 'Herr' (the German equivalent). Subject 3 would have preferred if the subtitles were a bit slower, and found the change in subtitles a little hectic. Subject 1 thought the quality of the subtitles was not very good and it would be too much effort to enjoy the film while using these subtitles.

**Table A-7:** German subtitles and English soundtrack: Lord of the Rings

| **Clip 7** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | ✓ |
| **Answer internal checks** | ✓✓ | ✓✓ | ✓✓ |
| **Soundtrack**<br>**Image**<br>**Subtitles** | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles |
| **Comprehensibility** | 6 | 4 | 5-6 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | ✓ | × |
| **Satisfaction** | 5 | 5 | 5 |

**Comments:** Subject 2 was not entirely satisfied with the subtitles, adding that the style was incorrect in places, using unusual translations and that the roughness of the speech in English was not transferred into the German subtitles. Subject 3 noted that background noises coupled with the strange accents of some of the characters made her use the subtitles more to understand the film.

**Table A-8:** German subtitles and Dutch soundtrack: Harry Potter and the Philosopher's Stone

| **Clip 8** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | ✓ | ✓ | Maybe |
| **Answer internal checks** | ✓✓ | ×✓ | ✓✓ |
| **Soundtrack Image Subtitles** | Image, Soundtrack, Subtitles | Soundtrack, Subtitles | Soundtrack, Subtitles |
| **Comprehensibility** | 5 | 6 | 4-5 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 6 | 6 | 6 |
| **Recommend subtitles for entire movie on DVD** | ✓ | ✓ | ✓ |

**Comments:** Subject 2 thought the translation of youth language was excellent and the style seemed very good. Subject 1 was not sure whether the youth language used was appropriate and thought it strange that the straight forward translations from Dutch into German were not used, and instead a new subtitle was created in German. Subject 3 thought the subtitles were too quick for children to read, but that the style was very appropriate. Even after these comments, all subjects thought the language of the subtitles was excellent.

**Table A-9:** German subtitles and English soundtrack: Chinatown

| **Clip 9** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | Maybe | ✓ | Maybe |
| **Answer internal checks** | ✓✓ | ×✓ | ×✓ |
| **Soundtrack Image Subtitles** | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles |
| **Comprehensibility** | 5 | 6 | 4 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | ✓ | × |
| **Satisfaction** | 4-5 | 6 | 4-5 |

**Comments:** All subjects said some parts of the dialogue during the telephone conversation were not subtitled and that some of the subtitles were shortened in strange ways. Overall they did not notice any obvious errors.

**Table A-10:** German subtitles and French soundtrack: Being John Malkovich

| **Clip 10** | **Subject 1** | **Subject 2** | **Subject 3** |
|---|---|---|---|
| **Understood Clip** | Maybe | ✓ | ✓ |
| **Answer internal checks** | ✓✓ | ✓✓ | ✓✓ |
| **Soundtrack Image Subtitles** | Image, Soundtrack, Subtitles | Image, Soundtrack, Subtitles | Image, Subtitles |
| **Comprehensibility** | 6 | 6 | 5 |
| **Acceptable for viewer who does not understand Soundtrack** | ✓ | ✓ | ✓ |
| **Errors** | × | × | × |
| **Satisfaction** | 6 | 5 | 6 |
| **Would watch entire movie with these subtitles** | ✓ | ✓ | ✓ |

**Comments:** No additional comments from any of the subjects.

# APPENDIX B

Dear Students

I am a PhD student conducting research in the area of audiovisual translation and I am currently designing a user evaluation of audiovisual material (in particular subtitles) which will take place during the second and third week in November. I am now looking for native speakers of German who would be willing to take part in this evaluation, and who have watched at least one movie on DVD with subtitles. The evaluation sessions would require no more than one hour of your time.

The aim of the session is to get your opinion on German subtitles provided on 6 DVD clips taken from a Harry Potter film. These subtitles are either selected, or "re-assembled" by a piece of software from a database of German subtitles created by human subtitlers in the normal fashion. Firstly I will ask you some questions relating to Harry Potter and subtitling in general. Then I will show you a clip on the TV and ask some questions after you have viewed the clip. This step will be repeated until you have viewed all 6 clips. The evaluation session will be conducted through English. We will be paying students 15 Euro for participating, and you will receive this in cash at the end of the session.

Kind Regards,

Marian Flanagan

# APPENDIX C

## CLIP 5

**Context: They are sitting at the dining table, and Harry notices Hermione is missing. He asks after her and Neville says she is in the girls' bathroom crying. Professor Quirrell comes running into the dining hall to tell everyone there's a troll in the dungeon. Harry and Ron must find Hermione before the troll gets to her….**

**HP: HARRY POTTER RW: RON WEASLEYHG: HERMIONE GRANGER PQ: PROFESSOR QUIRREL N: NEVILLE D: DUMBLEDORE P: PREFECT**

| | **Speaker** | **Subtitle** | | **Acceptable** |
|---|---|---|---|---|
| **5.en.1** | **HP** | Where's Hermione? | | |
| **5.en.2** | **N** | Parvati said she wouldn't come out of the bathroom. | | |
| **5.en.3** | **N** | She said that she'd been in there all afternoon, crying. | | |
| | | | **Translation 1** | |
| **5.en.4** | **PQ** | Troll in the dungeon! (no reps) | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |
| **5.en.5** | **PQ** | Troll in the dungeon! (no reps) | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Troll! Im Kerker! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |

| | | | | |
|---|---|---|---|---|
| **5.en.6** | **PQ** | Thought you ought to know (no reps) | Ich dachte, Ich sag Bescheid. | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Ich dachte, Ich sag Bescheid. | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Ich dachte, Ich sag Bescheid. | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |
| **5.en.7** | **D** | *Silence! | Seid Still! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Seid Still! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Seid Still! | Yes ☐ No ☐ Don't know ☐ |
| | | | | |
| **5.en.8** | **D** | Everyone will please not panic! | | |
| **5.en.9** | **D** | Now prefects will lead their house back to the dormitories | | |
| **5.en.10** | **D** | Teachers will follow me to the dungeons. | | |
| **5.en.11** | **P** | Gryffindors, keep up, please, and stay alert. | | |
| **5.en.12** | **HP** | How could a troll get in? | | |
| **5.en.13** | **RW** | Not on its own. | | |
| **5.en.14** | **RW** | Trolls are really stupid. | | |
| **5.en.15** | **RW** | Probably people playing jokes. | | |
| | | | **Translation 1** | |
| **5.en.16** | **RW** | *What? | Was? | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Was? | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Was? | Yes ☐ No ☐ Don't know ☐ |
| | | | | |

| | | | | |
|---|---|---|---|---|
| **5.en.17** | **HP** | Hermione! | | |
| **5.en.18** | **HP** | She doesn't know. | | |
| **5.en.19** | **RW** | I think the troll's left the dungeon. | | |
| **5.en.20** | **RW** | It's going into the girls' bathroom. | | |
| | | | **Translation 1** | |
| **5.en.21** | **HP** | Hermione, move! (no reps) | Hermine, komm raus da! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Hermine, komm raus da! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Hermine, beweg dich! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |
| **5.en.22** | **HG** | *Help! | Helfen! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |
| **5.en.23** | **HG** | *Help! | Helfen! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |
| **5.en.24** | **RW** | Hey, pea brain! | | |
| | | | **Translation 1** | |
| **5.en.25** | **HG** | *Help! | Helfen! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Hilfe! | Yes ☐ No ☐ Don't know ☐ |

| | | | | |
|---|---|---|---|---|
| | | | **Translation 1** | |
| **5.en.26** | **HP** | *Do something! | Tu was! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Etwas unternehmen! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Etwas! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 1** | |
| **5.en.27** | **RW** | *What? | Was? | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Was? | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Was? | Yes ☐ No ☐ Don't know ☐ |
| **5.en.28** | **HP** | Anything! | | |
| | | | **Translation 1** | |
| **5.en.29** | **HP** | *Hurry up! | Beeil Dich! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 2** | |
| | | | Beeil Dich! | Yes ☐ No ☐ Don't know ☐ |
| | | | **Translation 3** | |
| | | | Beeilt Euch! | Yes ☐ No ☐ Don't know ☐ |
| **5.en.30** | **HG** | Swish and flick | | |
| **5.en.31** | **RW** | Wingardium Leviosa | | |
| **5.en.32** | **RW** | Cool | | |
| **5.en.33** | **HG** | Is it dead? | | |
| **5.en.34** | **HP** | I don't think so. | | |
| **5.en.35** | **HP** | Just knocked out. | | |
| **5.en.36** | **HP** | Troll boogers. | | |

# APPENDIX D

## OPTIONS: CLIP 5

| **Subtitle** | **Options** | |
|---|---|---|
| **Silence!** | Ruhe! | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **Silence!** | Sei Still! | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **What?** | Was denn? | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **What?** | Und was | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |

| | | |
|---|---|---|
| **Help!** | Helft mir! | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **Do something!** | Tu doch was! | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **Hurry up!** | Beeilt euch. | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **Hurry up!** | Los, Beeilung! | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |
| **Hurry up!** | Los | Yes ☐ No ☐ Don't know ☐ |
| | | **Comments:** |

# APPENDIX E

**Background Information on the Subject:**

6. Gender: M F

7. Age group: 10-19 20-29 30-49 40+

8. Educational Background:
    a. High school diploma
    b. Undergraduate Degree
    c. Masters Degree
    d. Doctorate Degree
    e. Other:_____

**Background information relating to Harry Potter:**

9. Have you read any of the Harry Potter books?
    a. Yes
    b. No

10. How many?

11. If yes, what language did you read the Harry Potter books in?

    a. Language: _____

12. Have you seen any of the Harry Potter films?
    a. Yes
    b. No

13. How many?

14. If yes, what language did you see the films in?
    a. Language: _____

15. Did you watch any of the Harry Potter films with German subtitles?
    a. Yes
    b. No

16. How many? _____

17. If yes, why?
_____
_____

18. If no, why not?
_____
_____

19. Do you know the characters in the books/films?
    a. Yes
    b. No
    c. Not really – just the names, but not who they are

20. Would you consider yourself a fan of the Harry Potter series?
    a. Yes
    b. No
    c. Maybe

21. Do you understand spoken Dutch?
    a. Yes
    b. No

22. If yes, what level?
    a. Beginner
    b. Intermediate
    c. Advanced
    d. Fluent

**Background information relating to watching films on DVD with subtitles:**

23. How often would you watch films on DVD with subtitles?
    a. Once a week
    b. Every fortnight
    c. Once a month
    d. Twice a year
    e. Once a year
    f. Rarely
    g. Never

24. If you watch an English film on DVD, would you turn on the German subtitles?
    a. Yes
    b. No

25. If yes, why?
    _____
    _____
    _____

26. If no, why not?
    _____
    _____
    _____

27. If you are watching a film with German subtitles, and the soundtrack is in English, would you:

    a. Read the subtitles
    b. Just listen to the English soundtrack
    c. Or both?

28. If both,
    a. Are you comparing the English with the German subtitles
    b. Or just checking parts of the English you don't fully understand?

Comments:
_____

29. How often do you watch dubbed films?
    a. Once a week
    b. Every fortnight
    c. Once a month
    d. Twice a year
    e. Once a year
    f. Rarely
    g. Never

30. Do you enjoy watching dubbed films?
    a. Yes
    b. No

31. Why/Why not?
_____
_____
_____

32. Do you prefer watching
    a. Films dubbed into German
    b. Or English language films with German subtitles
    c. Or English language films with no German subtitles?

33. Why?
_____
_____
_____
_____

Information relating to Clip 1

1. After watching this clip, did you understand what was happening in the clip?
    a. Yes
    b. No

Comments:_____

2. In this clip, what did Ron say to Harry after Harry noticed Dumbledore had vanished from his card?

_____
_____

3. Did you understand the clip using a combination of:

    a. The soundtrack

    b. The image

    c. The subtitles,

4. Only two of these?

    a. Which ones?

    _____

5. Or only one of these?

    a. Which one?

_____
_____

6. On a scale 1-6 (6 being very comprehensible, 1 being incomprehensible) where would you locate the subtitles for this clip?

1 2 3 4 5 6

**Questions relating to the Subtitles:**

1. Are the following suitable?
    a. Font size: Yes/no – Comments _____
    b. Subtitle location: Yes/no – Comments _____
    c. Speed of subtitles: Yes/no – Comments _____

2. In your opinion are these subtitles acceptable to be used on a DVD, for people who would not understand the soundtrack?
    a) Yes
    b) No
    c) Maybe

    Comments: _____

3. On a scale of 1-6 (1 inappropriate and 6 very appropriate), how would you rate the style of the subtitles?

    1 2 3 4 5 6

Comments: _____

4. Did anything in the subtitles during this clip particularly:
    a. Bother you?
    b. Amuse you?
        a) _____
        b) _____

5. What subtitle errors, if any, did you notice in this clip?

_____
_____

6. On a scale of 1-6 (6 being not annoying at all, 1 being very annoying), where would you put the errors?

    1 2 3 4 5 6

Comments:_____

7. Did you notice any subtitles which seemed out of context?

_____

8. Do you remember any well-translated subtitles?

_____
_____
_____

**Specific to the Dutch soundtrack clips if the subject has no knowledge of Dutch:**

1. Focussing on the clips with the Dutch soundtrack, would you watch an entire movie with the same standard of German subtitles as those provided in these clips?

Why:_____
_____

Why not?

_____
_____

**After watching all the clips:**

1. Did you notice any subtitles which appeared a number of times throughout the 6 clips?
   a) Yes
   b) No

   Comments:_____

2. On a scale of 1-6 (6 being very satisfied, 1 being very dissatisfied), where would you rate the subtitles overall?

   1 2 3 4 5 6

   Additional comments on satisfaction/dissatisfaction:_____

3. In your opinion, were the subtitles of higher quality on the clips with the English language soundtrack or on the clips with the Dutch language soundtrack? Was there no difference?

   _____

# APPENDIX F

## PRE-VIEWING BRIEFING

This research looks at the reusability in new contexts of existing German translations for English movie subtitles. For instance, if the English subtitle 'Come on' has already been translated as 'Komm schon', we are interested in whether or not 'Komm schon' can be used to translate other instances of 'Come on' subsequently encountered in a particular movie or group of movies.

In this session:

- You will be presented with English subtitles for which a candidate translation into German already exists in a particular movie, as well as that candidate translation.

- You will be asked to give your opinion on whether or not the candidate translation is a good fit for the contexts in which we propose to use it. You can answer 'yes', 'no', or 'don't know'.

- You may be asked to give your opinion of alternative translations.

- You can make comments on any of the translated subtitles, or you can choose not to make any comments at all.

Any comments you do make will be of interest to us in a later analysis phase. The session will also be recorded on audio-cassette.

The results of this research will eventually be made available through a doctoral dissertation in DCU library. It is also intended to disseminate parts of this research in peer-reviewed publications. You will not be identified in any of these sources.

Many thanks for agreeing to participate in this study.

MARIAN

_____

I have been briefed on the aims of this study and I understand how the data I generate will be used. I give my full consent for these data to be used in Marian Flanagan's doctoral thesis.

_____

# APPENDIX G

## PRE-VIEWING BRIEFING

In this session, you will be shown a set of six movie clips taken from the first Harry Potter film, *Harry Potter and the Philosopher's Stone*. Each clip lasts approximately 2 minutes; although two clips are slightly longer (clips 4 and 6).

Three of the clips have an English soundtrack and three have a Dutch soundtrack. The session will either start with an English soundtrack clip or a Dutch soundtrack clip, and then alternate between the two languages for the following five clips.

All clips will be shown with German subtitles. These subtitles are either selected, or 're-assembled' by a piece of software from a database of German subtitles created by human subtitlers in the normal fashion.

The aim of the session is to get your opinion on these subtitles.

Firstly, you will be asked a number of background questions in relation to subtitling, your language ability, and your familiarity with the Harry Potter series.

Then you will be asked to view the first clip and to answer some more questions specifically about this clip. This step will be repeated until you have viewed all six clips.

The session will be recorded on an mp3 player, to capture any comments you make that are additional to those already recorded (in writing) by the researcher.

The results of this research will eventually be made available through a doctoral dissertation in DCU library. It is also intended to disseminate parts of this research in peer-reviewed publications. You will not be identified in any of these sources.

Many thanks for agreeing to participate in this study.

MARIAN May 2007 / NOVEMBER 2007

---

I have been briefed on the aims of this study and I understand how the data I generate will be used. I give my full consent for these data to be used in a doctoral dissertation at DCU, and in other relevant academic publications.

_____

# APPENDIX H

**Clip 1**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 10 | 10 | 10 |
| No | 7 | 7 | 7 |
| Don't know | 1 | 1 | 1 |

**Clip 2**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 13 | 16 | 16 |
| No | 5 | 2 | 2 |
| Don't know | 0 | 0 | 0 |

**Clip 3**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 17 | 15 | 14 |
| No | 5 | 7 | 8 |
| Don't know | 2 | 2 | 2 |

**Clip 4**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 11 | 11 | 11 |
| No | 3 | 3 | 3 |
| Don't know | 1 | 1 | 1 |

**Clip 5**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 20 | 26 | 23 |
| No | 15 | 9 | 12 |
| Don't know | 1 | 1 | 1 |

**Clip 6**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 6 | 4 | 5 |
| No | 3 | 5 | 4 |
| Don't know | 0 | 0 | 0 |

**Clip 7**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 17 | 17 | 21 |
| No | 6 | 6 | 3 |
| Don't know (2) | 1 | 1 | 0 |

**Clip 8**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 2 | 1 | 1 |
| No | 3 | 3 | 3 |
| Don't know | 1 | 2 | 2 |

**Clip 9**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 14 | 14 | 12 |
| No | 6 | 5 | 8 |
| Don't know | 1 | 2 | 1 |

**Clip 10**

| Response | HPCM (corpus AM) | HPCM + LOTRC (corpus BM) | HPCM + LOTRC +MGC (corpus CM) |
|---|---|---|---|
| Yes | 9 | 9 | 8 |
| No | 2 | 2 | 3 |
| Don't know | 1 | 1 | 1 |

# APPENDIX I

**Clip 1**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 | Subtitle 5 | Subtitle 6 |
|---|---|---|---|---|---|---|
| Yes | 3 | 4 | 1 | 2 | 3 | 3 |
| No | 3 | 2 | 3 | 7 | 0 | 9 |
| Don't know | 0 | 0 | 2 | 0 | 0 | 0 |

**Clip 2**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 |
|---|---|---|---|
| Yes | 9 | 6 | 3 |
| No | 9 | 9 | 3 |
| Don't know | 0 | 0 | 0 |

**Clip 3**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 | Subtitle 5 |
|---|---|---|---|---|---|
| Yes | 3 | 2 | 2 | 1 | 2 |
| No | 12 | 1 | 0 | 0 | 0 |
| Don't know | 0 | 0 | 1 | 2 | 1 |

**Clip 4**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 |
|---|---|---|---|
| Yes | 4 | 12 | 0 |
| No | 4 | 9 | 2 |
| Don't know | 1 | 0 | 1 |

**Clip 5**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 | Subtitle 5 |
|---|---|---|---|---|---|
| Yes | 3 | 2 | 3 | 3 | 0 |
| No | 3 | 3 | 0 | 0 | 8 |
| Don't know | 0 | 1 | 0 | 0 | 1 |

**Clip 6**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 |
|---|---|---|---|
| Yes | 3 | 10 | 9 |
| No | 3 | 15 | 9 |
| Don't know | 0 | 2 | 0 |

**Clip 7**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 |
|---|---|---|---|---|
| Yes | 5 | 3 | 0 | 3 |
| No | 4 | 3 | 9 | 3 |
| Don't know | 0 | 0 | 0 | 0 |

**Clip 8**

| Response | Subtitle 1 | Subtitle 2 |
|---|---|---|
| Yes | 4 | 2 |
| No | 6 | 17 |
| Don't know | 2 | 2 |

**Clip 9**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 |
|---|---|---|---|---|
| Yes | 2 | 3 | 2 | 11 |
| No | 0 | 0 | 0 | 24 |
| Don't know | 1 | 0 | 1 | 1 |

**Clip 10**

| Response | Subtitle 1 | Subtitle 2 | Subtitle 3 | Subtitle 4 |
|---|---|---|---|---|
| Yes | 3 | 9 | 2 | 2 |
| N | 0 | 1 | 1 | 6 |
| Don’t know | 0 | 2 | 0 | 1 |

# APPENDIX J

| Are there any subtitles which seem out of context? | | | | |
|---|---|---|---|---|
| | **Subtitle** | **Corpus A** | **Corpus B** | **Corpus C** |
| **Clip 1** | Ron holding his lunch and telling the lady with the trolley he didn't want to buy anything, he's all set! | ✓ | ✓ | ✗ |
| | Ron telling Harry to watch out | ✓ | ✗ | ✗ |
| | Harry greeting Ron as he comes into the carriage | ✓ | ✗ | ✗ |
| | Harry tells the lady with the trolley that he'll take the lot | ✓ | ✗ | ✗ |
| | The use of 'voll krass' for 'wicked' is inappropriate in this context | ✗ | ✗ | ✓ |
| **Clip 2** | Woman speaking about what the hat does | ✓ | ✓ | ✓ |
| | Verb 'to sort' into groups is incorrect | ✓ | ✗ | ✗ |
| | Hermione says sei unbesorgt, which is inappropriate in this context (too formal) | ✓ | ✗ | ✗ |
| | What the hat was saying | ✗ | ✓ | ✗ |
| | Harry says 'I'm good', incorrect translation | ✗ | ✗ | ✓ |
| **Clip 3** | Subtitle regarding how the stairs tend to move | ✓ | ✓ | ✗ |
| **Clip 4** | The verb 'to hang' was used in one of the subtitles, but this was totally out of context | ✓ | ✓ | ✓ |
| | At the beginning of the clip, one of the students says that Hermione is in the girls' bathroom, but the subtitle was very confusing | ✓ | ✗ | ✗ |
| | Troll not correction translation, and not appropriate in 'Troll im Kerker' | ✗ | ✓ | ✓ |
| | The use of 'gestorben' at the beginning of the clip, but nobody died in the clip | ✗ | ✓ | ✗ |

| | | | | |
|---|---|---|---|---|
| | The verb 'rufen' should not be used for 'to cry' | ✗ | ✓ | ✗ |
| | Harry and Ron's conversation later before the troll scene seemed peculiar | ✗ | ✗ | ✓ |
| | The use of maedenbadezimmer for girls' bathroom | ✗ | ✗ | ✓ |
| **Clip 5** | Hermione wonders what the flying keys mean, but the subtitle is not translated correctly | ✓ | ✓ | ✗ |
| | When Ron encourages Harry to get onto the broom and get the key they are looking for | ✓ | ✗ | ✗ |
| | At the beginning, Hermione says 'curios' when they walk into the room with all the flying keys – one subject suggests the translation was incorrect in this context | ✓ | ✗ | ✗ |
| | Hermione wonders what the flying keys mean, but the subtitle is not translated correctly | ✓ | ✗ | ✗ |
| **Clip 6** | The subtitle with Hermione saying being clever isn't the only important thing about being a good wizard | ✓ | ✗ | ✗ |
| | Hermione says I'll be ok to Harry instead of saying you'll be ok | ✓ | ✓ | ✓ |
| | Ron discussing the chess game at the beginning | ✓ | ✓ | ✓ |
| | Ron and Harry's conversation about chess pieces | ✗ | ✓ | ✗ |
| | **Total** | **18** | **13** | **9** |

# APPENDIX K

| Item that bothered the subjects | Corpus AM | Corpus BM | Corpus CM |
|---|---|---|---|
| Sometimes I could understand more of the Dutch soundtrack than the subtitles (comprehensibility) | ✓ | × | × |
| Translating Susan Bones' surname – *Susan Knochen* (style) | ✓ | × | × |
| *Kommt schon* to a group, but this is used to address one person (well-formedness) | ✓ | × | × |
| Harry says *eigenartig* when looking at the broom, he should probably use *komisch* (style) | ✓ | × | × |
| It didn't seem like a native speaker had translated the subtitles: incorrect style (style) | ✓ | × | × |
| Plural you was translated as *du* instead of *ihr* (well-formedness) | ✓ | × | × |
| Hermione used *Helfen* (infinitive) instead of *Hilfe* (imperative) (well-formedness) | ✓ | ✓ | × |
| in 100 years is a strange translation (comprehensibility) | ✓ | × | × |
| (Harry): *Es ist auch einfach* (It's too simple): Only got the meaning because of the English soundtrack (comprehensibility) | ✓ | × | × |
| Had to re-read quite a few of them (readability) | × | ✓ | × |
| (Hermione): *Ihr verbüßt ok, Harry. Du bist ein großer Zauberer. Wirklich.* (You'll be ok Harry. You're a great wizard, you really are). Inappropriate use of the verb *verbüßt* – register too high (style) | ✓ | × | × |
| The use of *to have* and *to be* are mixed up (well-formedness) | × | ✓ | × |
| Words that don't exist: *Eulerei* (owlery) and *wutschen und wedeln* (swish | ✓✓ | ✓ | ✓ |

| | | | |
|---|---|---|---|
| and flick) (well-formedness) | | | |
| Simply reading 'normal words' is problematic (comprehension) | × | × | ✓ |
| (Hermione): *Und Harry, es sei vorsichtig*. (And Harry just be careful). No need for the pronoun *es* | × | × | ✓ |

# APPENDIX L

| Item that amused the subjects | Corpus A | Corpus B | Corpus C |
|---|---|---|---|
| (Ron): *Wir wollen ein überzeugter, altmodischen, vermutlich rostig* | ✓ | × | × |
| (Hermione): *Bücher und Schlau-sein, es gibt mehr wichtige Dinge* | ✓ | × | × |
| The appropriate use of kids' speak | ✓ | × | × |
| (Hermione): *Ok, sei unbesorgt* - Not idiomatic in this context | ✓ | × | × |
| (Ron): *Harry, was ist das?* - Not idiomatic in this context | ✓ | × | × |
| No correct subtitles in full | ✓ | × | × |
| *wahrlich* mentioned at the beginning of the clip | ✓ | × | × |
| (Ron): *Das macht Dinge ein bisschen schwierig* - dry humour | ✓ | × | × |
| Some words were out of context and it made it funny to read | × | ✓ | × |
| The misuse of the language | × | ✓ | × |
| Translating Susan Bones' surname – *Susan Knochen* | × | ✓ | × |
| I could imagine in many instances what the subtitle should have said! | × | ✓ | × |
| The subtitles describing the pictures by the stairs were good | × | ✓ | × |
| Some parts seemed funny because of the errors | × | × | ✓ |
| The subtitles didn't capture the mood of the conversation | × | × | ✓ |

# APPENDIX M

## Well-formedness Codes: Corpus AM

| **Linguistic issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Subtitles contain too many errors | 8 |
| Incorrect grammar | 6 |
| Quality of German translation is poor/incorrect translations | 3 |
| Incorrect word order | 2 |
| Incorrect conjugation of verbs | 1 |
| Too many incorrect words | 1 |
| Some subtitles contain words that are not used in the German language | 1 |
| Terminological errors | 1 |

| **Comprehension issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Movie clips contain incomprehensible subtitles | 8 |
| Subtitles are too confusing | 2 |
| Without prior knowledge subtitles are incomprehensible | 1 |
| Subtitles are too fast which inhibits comprehension | 1 |

| **Subtitles acceptable with improvements** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The subtitles are acceptable for use on a DVD, but the complicated subtitles are incorrect and they need improving | 4 |
| The subtitles might be of some help, but at the moment they are too confusing and distracting to concentrate on other things happening in the clip | 2 |
| Subtitles are comprehensible and could be used but the quality of German is bad | 1 |

| **Subtitles acceptable** | |
|---|---|
| **Codes** | **Number of references within the code** |
| If the viewer is interested in the movie the could use the subtitles | 1 |
| The short subtitles were translated very well and were very appropriate | 1 |

# Well-formedness Codes: Corpus BM

| **Linguistic issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Quality of German translation is poor | 5 |
| Subtitles contain too many errors | 5 |
| Incorrect grammar | 4 |
| Subtitles do not express the meaning correctly | 3 |
| The subtitles appear to be translated using a free-online MT system | 1 |
| Unnatural direct or word-for-word translations | 1 |
| Subtitles contain awkward expressions | 1 |
| Incorrect word order | 1 |

| **Comprehension issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Movie clips contain incomprehensible subtitles | 8 |
| Without prior knowledge, a viewer wouldn't understand the subtitles | 1 |
| Subtitles are too fast which inhibits comprehension | 1 |

| **Subtitles comprehensible, but not acceptable** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The subtitles were helpful, but not perfect | 3 |
| The subtitles would give the viewer a hint of what the movie is about, but they still contain errors | 1 |

| **Lack of enjoyment** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The subtitles are suitable for comprehension, but not useful for viewers to enjoy the movie | 3 |

| **Subtitles acceptable with improvements** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Not exactly acceptable, but a viewer could watch a movie on DVD using these subtitles if there was no other option | 3 |
| The subtitles would be suitable for a learner of German, but not for a native speaker | 2 |
| The subtitles might be suitable, but they contain some grammar mistakes | 2 |
| The subtitles are acceptable 80% of the time, but need improvements for the 20% of mistakes | 1 |
| The subtitles are acceptable if freely available to download from the Internet, but they would not be good enough for a DVD | 1 |
| The subtitles are understandable, so they might be suitable for use on DVD | 1 |
| The subtitles are a mix of good and bad | 1 |

| **Subtitles acceptable** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The storyline is understandable using these subtitles | 1 |

# Well-formedness Codes: Corpus CM

| **Linguistic issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Subtitles contain too many errors | 15 |
| Incorrect grammar | 5 |
| Subtitles do not express the meaning correctly | 4 |
| Quality of German translation is poor | 2 |
| Unnatural direct or word-for-word translations | 1 |

| **Comprehension issues** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Movie clips contain incomprehensible subtitles | 11 |
| Subtitles are lacking sense | 7 |
| Subtitles are too confusing | 1 |
| Without prior knowledge subtitles are incomprehensible | 1 |
| The longer subtitles are incomprehensible | 1 |
| Movie clip is only comprehensible when using the image and soundtrack | 1 |

| **Subtitles comprehensible, but not acceptable** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The subtitles were helpful, but not perfect | 5 |

| **Lack of enjoyment** | |
|---|---|
| **Codes** | **Number of references within the code** |
| Watching subtitles with so many errors is annoying and not enjoyable | 3 |
| Subjects would expect better quality subtitles, especially when the soundtrack language is unknown to the viewer, as a lot of information is then lost | 1 |
| The subtitles are comprehensible, but not acceptable | 1 |
| The subtitles are suitable for comprehension, but not useful for viewers to enjoy the movie | 1 |

| **Subtitles acceptable with improvements** | |
|---|---|
| **Codes** | **Number of references within the code** |
| The subtitles are not exactly acceptable, but a viewer could watch a movie on DVD using these subtitles if there was no other option | 2 |
| The subtitles are understandable, so they might be suitable for use on DVD | 1 |
| The subtitles contributed to understanding, but are not acceptable for an entire DVD | 1 |
| Apart from a few bad quality subtitles, subjects would use these subtitles | 1 |

# APPENDIX N

| Are there any well-translated subtitles in the movie clips? | | | | |
|---|---|---|---|---|
| | **Subtitle** | **Corpus A** | **Corpus B** | **Corpus C** |
| **Clip 1** | Wicked (*) | ✓ | ✓ | ✓ |
| | Flavours of sweets | ✓ | ✓ | ✓ |
| | Introductions between Harry and Ron | ✓ | ✓ | ✓ |
| | He's gone | × | ✓ | ✓ |
| | You can't expect him to hang around all day, can you? | × | ✓ | × |
| **Clip 2** | The sorting hat deciding on the houses | ✓ | ✓ | × |
| | Ron talking about Hermione | ✓ | ✓ | ✓ |
| | Proper names | ✓ | × | ✓ |
| | Harry touches his head and says he's fine | × | ✓ | |
| **Clip 3** | The school children talking about the pictures along the stairs | ✓ | × | ✓ |
| | Prefects telling the younger students to keep up and come along (**) | ✓ | ✓ | ✓ |
| | Prefect welcoming the students to the Gryffindor common room | ✓ | ✓ | ✓ |
| | Prefect telling the younger students to 'follow me' | ✓ | ✓ | ✓ |
| | The picture asking for the password to the common room | ✓ | ✓ | × |
| | Prefect showing the students the way to the common room | ✓ | ✓ | ✓ |
| **Clip 4** | The professor coming into the dining hall telling everyone about the troll in the dungeon | × | ✓ | ✓ |

| | | | | |
|---|---|---|---|---|
| | Ron calling the troll ‘pea brain' | ✓ | ✓ | ✓ |
| | Ron telling Hermione to come out of the toilet cubicle | × | ✓ | ✓ |
| | Many of the short subtitles during the Troll scene in the girls’ bathroom | ✓ | ✓ | ✓ |
| **Clip 5** | Short subtitles were translated correctly | ✓ | ✓ | ✓ |
| | The description of the key with the broken wing | × | × | ✓ |
| | Harry ordering Ron and Hermione to ‘catch the key' and to ‘hurry up’ | ✓ | ✓ | ✓ |
| | Hermione comments on the strange situation they find themselves in(*) | × | × | ✓ |
| **Clip 6** | The chess-related terminology | ✓ | ✓ | ✓ |
| | Short subtitles including 'wait a minute', I have to go’, ‘there are more important things', ‘you’re a great wizard Harry’, ‘don’t move, we’re still playing’, and ‘not me, not Hermione, but you’. | ✓ | ✓ | ✓ |
| **Total** | | **18** | **21** | **21** |

# APPENDIX O

## Corpus AM

| **Helpful subtitles (6)** |
|---|
| The subtitles are helpful to give you some idea if you don't understand the soundtrack. |
| Short subtitles were translated well and the grammar was correct. |
| Sometimes very short idiomatic subtitles were correct. |
| The subtitles might be helpful sometimes. |
| You can understand some information from the subtitles but it's important that the grammar is correct. |
| There was one clip with good grammar. |

| **Unsuitable subtitles for a commercial DVD (2)** |
|---|
| The quality of the subtitles is very bad and they should not be used on a DVD. |
| The subtitles are useful for understanding when you don't have knowledge of the soundtrack, but would be very disappointed if they were on a DVD or in the cinema. |

| **Bad quality subtitles (3)** |
|---|
| In general the grammar was poor. |
| Incorrect grammar is very annoying. |
| The subtitles are of bad quality. |

| **Machine Translation Technology and Post-editing (5)** |
|---|
| It is difficult for the MT system to translation ambiguous words. |
| When a word has two meanings, the wrong meaning was often chosen. |
| I think a bad human translation is better than a translation done by a computer. |
| The translation by a computer just seems to be a word-for-word translation. |
| You can tell it's a machine translation - not human-like. |

| **Learners of German (2)** |
|---|
| Children would have many problems understanding these subtitles. |
| If people used these subtitles they would learn incorrect German. |

| **Using prior knowledge (PK) to understand the movie (2)** |
|---|
| I used PK to understand what was happening. |
| I used PK to understand all of the clips. |

| **Unsatisfactory subtitles (8)** |
|---|
| I kind of understand from the subtitles, but reading these subtitles is exhausting. |
| I have to re-read the subtitles all the time which is annoying. |
| I had to re-read many subtitles and re-structure them myself. |
| I'm not happy at all with the translations, very difficult to follow, and I am not familiar with the subject matter. |
| I am relying a lot on image or soundtrack to understand what is happening. |
| There is a lot going on and it's difficult to concentrate on the subtitles when they contain errors. |

| |
|---|
| You can't use subtitles with errors, as they are very annoying to read. |

**Dubbing versus Subtitling - 0**

## Corpus BM

| **Unsuitable subtitles for a commercial DVD (5)** |
|---|
| Some subtitles helped, but you couldn't use them on a purchased DVD, poor quality, clip 6 was the worst. |
| Main comment, mistakes in time (plural/sing), sentence construction, understandable, but wouldn't purchase a DVD with this quality, not satisfying, wouldn't enjoy watching a film with these subtitles. |
| Such subtitles shouldn't be used, even on a DVD. |
| I wouldn't use them on the English version, if you get it for free, then ok, but they should be of better quality if you are paying for them. |
| English structure, directly translated, quality not good enough for a purchased DVD. |

| **Bad quality subtitles (5)** |
|---|
| To answer some of the questions I used prior knowledge from books, I'm annoyed the subtitles are of such low quality, most of them are word for word translations, most annoying |
| Quality not great, but understandable, depends on what you expect from subtitles, they are ok, but they should be correct all the time |
| The quality of the subtitles is pretty poor - they are machine translated, not human translations. |
| The quality of the subtitles is very bad – I wouldn't use them |
| Really surprised subtitles are so bad, usually good standard of subtitles |

| **Learners of German (1)** |
|---|
| I wouldn't use these subtitles as they do not give a good impression of German subtitles, especially showing them to children or learners of German |

| **Unsatisfactory subtitles (6)** |
|---|
| Still possible to understand the film and still be able to enjoy it, but lots of details are unsatisfactory, I can think of quite a few better translations, expressions weren't too difficult and it's exhausting if you have to re-read subtitles. |
| Translation of words not bad, grammar terrible. |
| The subtitles should be part of the entertainment, not exhausting to watch a film. Watching a film only in parts is difficult. |
| It would be better to watch the movie just listening to the soundtrack. |
| I would only watch famous films or very interesting films with subtitles, and I would prefer German language films or English soundtrack. |
| These subtitles are harder to understand vs. German or English human subtitles. |

**Using prior knowledge (PK) to understand the movie - 0**
**Helpful subtitles - 0**

**Dubbing versus Subtitling - 0**
**Machine Translation Technology and Post-editing - 0**

## Corpus CM

| **Helpful subtitles (4)** |
|---|
| The short subtitles were translated well. |
| The second English clip is almost acceptable. |
| Amazing it can translate that well, not perfect, but understandable. |
| If I had no other choice, I would use these subtitles to watch a film. |

| **Machine Translation Technology and Post-editing (3)** |
|---|
| I can see where the problems occur most and perhaps post-editing could be a possibility |
| For post editing purposes, these subtitles would be useful, as the post-editor can change and rearrange the language as it's supposed to be used and it might make the job easier and reduce unnecessary pressures |
| If all the MT output is wrong, the subtitler could start again from scratch |

| **Unsatisfactory subtitles (3)** |
|---|
| I understood the subtitles, but it is annoying that the wrong tenses were used and wrong word order. |
| I could understand the English clips with no prior knowledge, and I could not understand the Dutch clips with no prior knowledge. |
| They were useful for informing the viewer, but not useful for enjoying the film. |

| **Dubbing versus Subtitling (1)** |
|---|
| Dubbed films are always correct in terms of meaning and easy to understand. |

| **Unsuitable subtitles for a commercial DVD (4)** |
|---|
| Could never use these subtitles on a DVD. |
| I understand the main meaning but I want to understand everything – I'm not used to only understanding 3-5 subtitles out of ten. |
| I wouldn't use them on a DVD as there are too many mistakes. |
| Some subtitles were ok, but others were very bad and none were good enough for purchased DVD. |

| **Bad quality subtitles (1)** |
|---|
| The subtitles shouldn't be translated; I would prefer to listen to the soundtrack |

| **Using prior knowledge (PK) to understand the movie (2)** |
|---|
| Some subtitles were ok - mostly the short ones were translated correctly, but the longer ones were messed up, most of the time I was using prior knowledge to understand. |
| I used prior knowledge to understand the storyline in many cases. |

**Learners of German - 0**

# APPENDIX P

**Corpus AM**

**YES**

| **Subtitles would be useful** |
|---|
| The quality of the German translation wasn't the best at times, but I could use them to follow the action |
| The subtitles are better than nothing, but the quality would also bother the viewer as sometimes they were basically quite funny and distracting |
| I would use them - I have the context from the image, but I would also have to put up with a lot of errors |

| **Helpful subtitles** |
|---|
| The subtitles were helpful |
| I understand the movie and what's going on |
| I could use these subtitles |
| I would use them if I really had to |

**MAYBE**

| **No alternative option** |
|---|
| Borderline - I probably would use them if I didn't understand the soundtrack |
| I would watch a movie with these subtitles if there was no other option, as they are better than nothing |
| I could ignore some subtitles, and use the others |
| If I had no choice, I can get a gist from these subtitles |
| If I had no other choice I would use the subtitles |
| I would try watching them and then turn them off if they became really annoying |
| If I was interested in the film I would use them, no otherwise, as the speed of the subtitles is too fast and I couldn't read many of them |
| If I liked the film, then I would watch it with these subtitles, but otherwise it would be too difficult and not enjoyable |
| I might use them in the beginning, switch them on occasionally, but I would need some prior knowledge to fully understand them |
| I would use them but only in an emergency, as it is difficult to follow the film using the subtitles |
| If you had to use them to get the gist, fine, but otherwise I wouldn't turn them on |
| If there was no other option I would use them to get the gist of the movie, but I would really prefer not to |
| I could understand a few, got a very small idea of what was happening, so suppose I would |
| I would just ignore any bad subtitles and watch them for the comedy factor |
| It was difficult at times to understand, and therefore I might miss one or two trying to decipher the translation |

**NO**

| **Quality too bad** |
|---|
| Subtitles are too bad, can't follow story |
| No sense |
| Didn't understand anything |
| Incomprehensible |
| Incomprehensible |
| Couldn't understand anything using the subtitles |

| **Use image and soundtrack** |
|---|
| I would turn off the subtitles and just listen to the soundtrack & watch image; |
| I wouldn't watch the film with the subtitles; I would try and understand the Dutch; |
| I would just turn off the subtitles and listen to the Dutch soundtrack |
| I wouldn't use them and I would try to understand the Dutch |

| **Unsatisfactory and Annoying** |
|---|
| It would be very annoying to watch a whole film with these subtitles |
| I wouldn't use them or watch the film with them |
| too difficult to use these subtitles for 2 hours, it's ok for 10 minutes |
| if I didn't know the story in advance, it would be too confusing to understand |
| The subtitles are too annoying |
| I would leave them out, very confusing |
| The subtitles are too difficult/exhausting to read |
| After a while the mistakes would be very annoying |
| I wouldn't enjoy the movie – I can get the meaning but no sense of what is going on |
| If I had no choice I would use them, but this is unlikely |
| I would read the book, it would be more fun |

**Corpus BM**

**YES**

| **Helpful subtitles** |
|---|
| I would understand using the subtitles and images |
| The subtitles are good enough |
| I could use the subtitles to understand the movie |
| I would use them |
| I would use these subtitles to watch the film |

**MAYBE**

| **No alternative option** |
|---|
| I don't really want to, but if there was no other option I would use them on my own, but definitely not in a group |
| If I really wanted to see the movie, I would use them, but not if I wasn't too interested in the movie |
| If there was no other choice, but then I wouldn't understand everything using subtitles of this quality |
| Maybe, it really depends on the alternatives |
| I would watch a movie with these subtitles, if there was no other option |
| If I really wanted to see this movie, I would use the subtitles |
| If there was no other option, I would use these subtitles |
| If I had to, then I would use them |

**NO**

| **Unsatisfactory and Annoying** |
|---|
| The mistakes are annoying and not enjoyable to watch. |
| You end up concentrating too much on mistakes |
| The subtitles are conflicting when reading German and listening to English (known language) |
| The subtitles are too mixed up, confusing and annoying |
| The mistakes are too annoying, so I would use another option preferably |
| The subtitles are not enjoyable to read, perhaps use them as a reference, but couldn't use them to enjoy and understand the film |
| Subtitles are too annoying |
| I couldn't use these word-for-word translations as subtitles |
| The movie wouldn't be enjoyable using these subtitles |
| It's annoying always having to think about the subtitles |

| **Comprehension difficulties** |
|---|
| I wouldn't understand the movie using these subtitles |
| As I have no prior knowledge of movie, it would be very difficult to follow with these subtitles |
| If you had no prior knowledge, you would not understand what was going on |
| I couldn't use the subtitles to understand the movie |
| The subtitles are too annoying, too mixed up and not understandable |
| I would have problems understanding the subtitles |
| I wouldn't use the subtitles; I tried to make out the meaning from the Dutch (unknown language) |

| **Subtitles require too much effort** |
|---|
| I probably wouldn't use these subtitles as it is too exhausting to read them |
| I don't think so, as I wouldn't understand and it would be annoying - too much effort required |
| I would have to re-read them and it's not enjoyable |
| The subtitles require me to think about what they really mean |
| The subtitles are too hard to understand and not enjoyable |
| Reading the subtitles is not enjoyable |
| The subtitles are too difficult to read and not enjoyable |
| I wouldn't use these subtitles as the grammar is so bad and I would spend too long thinking about the meaning |
| If you didn't know the film, too difficult to understand |

| **Quality too bad and unhelpful** |
|---|
| The subtitles are too bad and the film is not interesting as a result |
| I would wait until I had a chance to watch the film with a different language combination, perhaps German dubbed |
| The subtitles contain too many mistakes |
| I was concentrating more on the images than using the subtitles |

**Corpus CM**

**YES**

| **Helpful and satisfactory subtitles** |
|---|
| I would use them because I could understand what was meant |
| I would watch it with the subtitles provided |
| The image and subtitles give you meaning, so if it's just one or two subtitles that are bad quality, who cares?! |
| I could understand the story using the subtitles |
| I could understand from these correct subtitles and was not distracted when watching film |
| I would use these subtitles |
| I could get most of the meaning |

**MAYBE**

| **Viewer very interested in movie** |
|---|
| If I really wanted to watch the film I would use the subtitles, but otherwise it depends on movie genre |

| **Viewer with prior knowledge** |
|---|
| If I had read the book in advance it would be ok, but if I didn't know the story, I wouldn't use them |

| **Viewer with little interest in movie** |
|---|
| I might watch the movie with the subtitles if I wasn't that interested in the film, but I would stop if I have to re-read subtitles and think about the meaning as it would be very annoying |

| **Depends on the movie** |
|---|
| It depends on the movie whether or not to use the subtitles; it was difficult to understand at times, so I might stop watching after a while |

| **No alternative option** |
|---|
| I would use them as it's better than having none |
| I probably would use the subtitles |

**NO**

| **Comprehension difficulties** |
|---|
| I don't think so as comprehension is a problem and I wouldn't enjoy the movie |
| There's no sense in watching a movie you can't understand |
| I couldn't understand what was happening |
| I wouldn't understand the movie |
| It is difficult to follow the subtitles |
| I wouldn't understand what was happening |
| I couldn't understand |
| I couldn't understand the movie |
| The subtitles are really annoying and I can't understand |
| I couldn't understand the story |
| The subtitles are incomprehensible |
| I didn't understand anything |
| I couldn't understand fully |
| There are too many mistakes and I couldn't enjoy or understand the movie |
| I couldn't understand what was happening |
| There are too many errors making it complicated to understand |

| **Unsatisfactory and Annoying** |
|---|
| I wouldn't understand the movie and can get only some meaning, which is annoying to watch |
| The subtitles wouldn't please the viewer |
| I wouldn't enjoy the movie |
| You could understand the subtitles but it wasn't enjoyable |

| **Subtitles require too much effort** |
|---|
| I would probably just turn off subtitles and can guess what is happening because it's Dutch (unknown language) |
| I would prefer to try to use the soundtrack as I just couldn't use the subtitles |

| **Quality too bad and unhelpful** |
|---|
| The correct translation was not provided |

| **Viewer very interested in movie** |
|---|
| I wouldn't watch it if I really wanted to see the film, as I can't spend the whole time wondering what is happening |