FACULTY OF ARTS AND PHILOSOPHY

ESTIMATING TRANSLATION DIFFICULTY BASED ON READABILITY SCORES, SUBJECTIVE EVALUATION AND PROCESS DATA

Word count: 18,247

Lise Verstraete
Student number: 01610305

Supervisors: Prof. Dr. Lieve Macken, Bram Vanroy

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Arts in Translation (Dutch, English, Spanish)

Academic year: 2019 – 2020

GHENT UNIVERSITY

*** Verklaring i.v.m. auteursrecht**

Preambule

Door de coronamaatregelen hebben we het onderzoek voor de masterproef gedeeltelijk moeten aanpassen.

Het originele onderzoek bestond uit twee stappen. Eerst werden twee Engelse teksten opgestuurd naar vertaaldocenten en werd hen gevraagd om een vragenlijst in te vullen in verband met de teksten en om moeilijke items in de teksten aan te duiden. Als tweede stap waren we van plan om studenten uit de master Vertalen en uit het postgraduaat Computer-Assissted Language Mediation (CALM) twee teksten te laten vertalen en tijdens het vertalen hun toetsenbord te loggen met Translog en ook een eye tracker te gebruiken om de oogbewegingen te tracken. Het vertalen met de eye tracker zou gebeuren op computers van de UGent. De eye tracker zou net getest worden toen de coronamaatregelen in werking traden.

Omwille van de coronacrisis hebben we daarom geen data kunnen verzamelen over het vertaalproces met behulp van de eye tracker, aangezien niemand nog naar de universiteitsgebouwen mocht gaan. Samen met de promotor hebben we toen besloten om de studenten hun eigen laptop te laten gebruiken voor het vertalen, maar dan zonder eye tracker. Alle studenten moesten dan wel eerst Translog op hun laptop installeren.

Ook heb ik vertraging opgelopen met de dataverzameling van de docenten. De coronamaatregelen zorgden er namelijk voor dat de docenten extra veel werk hadden omdat ze hun lessen moest aanpassen om digitaal les te geven, examens moesten aanpassen, enz. Omdat ik de docenten een week voor de lockdown had aangeschreven, zijn er verschillende mijn e-mail wat vergeten door de extra drukte, waardoor het langer heeft geduurd om de nodige informatie te verzamelen.

Deze preambule werd in overleg tussen de student en de promotor opgesteld en door beide goedgekeurd.

Acknowledgments

This master thesis would not have been possible without the contribution of several people and I acknowledge all of them for their involvement.

First of all I would like to thank my promotor, Prof. Dr. Lieve Macken, and my co-promotor, Mr Bram Vanroy. Both of them enthusiastically encouraged my work and provided a lot of guidance, many ideas and practical support. I have had many virtual and real meetings with them and each time I learned a lot. Their valuable comments and suggestions on the draft texts also substantially influenced the final document.

I also would like to thank all the participants in the experiment: 9 teachers and 10 students. Their contribution was essential to this master thesis and is appreciated a lot.

Finally, I would also like to thank my parents for their stimulating support throughout my studies and for teaching me the art of hard work.

Table of Contents

Abstract

In the field of textual translation, both training and professional practice, it is important to know what makes a text difficult to translate. However, there is only limited research done on this topic and there is not yet any consensus on which textual elements cause difficulty for translation or how to estimate the translation difficulty of a text. In this master thesis we have conducted an experiment with the aim of evaluating the relationship between the translatability of two texts and three different types of data: readability scores, subjective evaluation and process data. The texts for the experiment were selected based on their readability scores: one easy and one difficult text. The participants consisted of two groups, teachers of English translation and translation students. They were asked to rate the translatability of both texts and mark items that they considered difficult to translate. The students were also asked to translate the texts while their keyboard activity was being logged. We conclude that readability is a good indicator of translatability. The correlation with respect to difficult items marked by teachers and students, however, is weak. For the process data, we focused on the duration and number of edits, and found that they are reliable indicators of translation difficulty. We conclude by suggesting some topics for future work. (216 words)

Abbreviations

CNA	Choice Network Analysis
CRITT	Centre for Research and Innovation in Translation and Translation Technology
Dur	Duration
Dxx	teacher x
IQR	Inter-Quartile Range
LexTALE	Lexical Test for Advanced Learners of English
Nedit	Number of edits
Pxx	participant x
T0x	text x
TPR-DB	Translation Process Database

List of tables

List of figures

1. Introduction

In the field of textual translation, both training and professional practice, it is useful to be able to estimate translation difficulty of a text. Hence the question: what makes a text difficult to translate? There is some, but only limited research available on this topic. There is also no consensus yet about which textual elements cause difficulty for translation. This master thesis addresses this problem of estimating translation difficulty based on readability scores, subjective evaluation and process data.

We begin by giving background information about the important concepts: what is readability and translatability, what research has been carried out in this area, and what are the findings? What are the textual indicators of readability and translatability? We also discuss which features have been suggested to estimate translatability, and which have been proven to be applicable or not relevant and why. Next, we give a short overview of the CRITT Translation Process Database (CRITT, s.d.), which is publicly available and has been used in many different studies. This discussion can be found in section 2 and is followed by our research objective in section 3.

In section 4, we sketch the experiment that we organized with the aim of finding an answer to our research questions. Two texts were selected for the experiment based on their readability scores: one simple and one difficult text. The participants in the experiment were both experienced teachers and translation students. They were asked to rate the translatability of both texts and to mark textual elements that they considered difficult to translate. The students were then also asked to translate the texts and we logged their keyboard activity during the translation.

The body of the master thesis presents all the data that results from this experiment: the subjective ratings of the overall translation difficulty of each text, the difficult items on segment or item level, and the process data collected during the translation itself, specifically duration and number of revisions. In light of these data, we discuss the relationships between readability and translatability, subjective rating of difficulties by teachers versus students, and the correlation between the two process features and between the process features and the subjective ratings.

We make some suggestions for future research in section 7. In section 8, we conclude that readability can be an indicator of translatability when two texts with a big difference in readability scores are compared. However, the correlation with respect to difficult items marked by teachers and students is weak. For the process data, we found that duration and number of revisions are indeed reliable indicators of translation difficulty.

2. Background

This section gives a summary of the background information about translation difficulty: what are the different concepts involved with translatability and what research has been conducted so far? We first describe what readability is, how it can be predicted and which indicators are important. Next, we discuss the concept of translatability, the possible indicators and then we give a short overview of the research. Finally, we discuss the CRITT Translation Process Database (CRITT, s.d.), a database that was used in our experiment.

2.1. Readability

2.1.1. Definition

Readability does not have one conclusive definition, because researchers use different definitions in their studies, often depending on the purpose of their study. Dale & Chall (1949) define readability as follows:

> Readability is the sum total (including the interactions) of all those elements within a given piece of printed material that affects the success that a group of readers have with it. The success is the extent to which they understand it, read it at an optimum speed, and find it interesting. (Dale & Chall, 1949, p. 23)

They state that the readability of a text has an influence on the understanding of the text, the reading speed and the interest in the text. According to Dale & Chall (1949) there is not one element that defines readability, but multiple elements, which also interact with each other. The developer of the SMOG (Simple Measure of Gobbledygook) Grade, McLaughlin (1974) uses the following definition for readability: “Readability is generally taken to mean that quality of written material which induces a reader to go on reading” (p. 367). McLaughlin gives a general definition of readability, without going into specifics. The definition does not give an idea about what influences the desire of the reader to continue reading, similar to the definition of Dale & Chall. According to DuBay (2004), “readability is what makes some texts easier to read than others” (p. 3). Much like McLaughlin’s definition, DuBay’s definition focuses on the general meaning of readability. The difference between McLaughlin’s and DuBay’s definition is that DuBay mentions that readability has to do with how difficult a text is, whereas McLaughlin approaches readability from the perspective of the reader and how likely the reader is to continue reading. DuBay also warns that readability is often confused with legibility. However, these words do not mean the same thing, as legibility is more concerned with typeface and layout. Finally, Jensen (2009) defines readability as “the ease with which the text is likely to be read and comprehended” (p. 63). Similar to Dale & Chall’s definition, Jensen highlights that the readability

of a text not only impacts how fast a reader reads a text, but also how well the reader understands what the text is about.

In this master thesis, 'readability' is defined as all the elements that have an influence on how well a reader will understand a text, how fast he will read the text and how interested he is in the text. Readability is determined by more than just textual characteristics. Characteristics of the reader also influence how easy a text can be understood. These reader characteristics can for example be the background, the education level of the reader or the interests of the reader. It is therefore important to take also these characteristics into account when trying to assess the readability of a text.

2.1.2. Predicting readability

Readability has traditionally been predicted with readability formulas. A readability formula is "a mathematical formula, typically consisting of a number of variables (i.e. text characteristics) and constant weights, intended to grasp the difficulty of a text." (De Clercq et al., 2014, pp. 295-296). These formulas only use textual elements to predict readability. The two variables that are most often used in these formulas are vocabulary difficulty and average sentence length, because these variables give the most accurate prediction of the difficulty of a text according to human evaluation (DuBay, 2004).

Some of the most commonly used readability formulas are the following: Flesch-Kincaid Readability Test, Flesch Reading Ease Formula, Dale-Chall Formula, Gunning Fog Index, SMOG Formula and Fry Readability Formula (DuBay, 2004). These formulas are distinct from each other in three ways: number of variables, the weight of the variables and the interpretation of the score. Some examples of variables that are used in readability formulas are average sentence length, the number of polysyllabic words and number of syllables per word (Flesch, 1948; McLaughlin, 1969). How many variables are included in a formula varies and some formulas even only contain one variable. The SMOG Formula is an example of a formula that uses only one variable, namely the number of polysyllabic words, whereas the Flesch Reading Ease formula, for example, contains two different variables. In each formula the weights that are attributed to the variables are different, resulting in different readability scores. The interpretation of the score is also different for some formulas. For most formulas, such as the Dale-Chall Reading Grade Formula, the rule is the higher the score, the more difficult the text is. The interpretation is opposite for the Flesch Reading Ease Formula, where a high score is indicative of an easy text (DuBay, 2004).

The above-mentioned traditional formulas have been developed in the early days of readability research. Since the early 20th century researchers have been conducting research into readability and that is when they began developing the first readability formulas (Sun, 2015). The formulas have already known a long tradition and their success stems from two aspects. Firstly, the formulas are easy

to work with and they make it easy to calculate readability, which was especially important when researchers manually calculated readability scores. The reason why they are easy to work with is that these formulas are based on mathematical measures that can be easily extracted from the text. Additionally, most traditional readability formulas do not use too many variables. Using more variables makes the formula more complex and difficult to work with, while there is no big difference in accuracy (DuBay, 2004). Secondly, these formulas are also widely available. The Flesch-Kincaid Grade Level formula, for example, is used in Microsoft Word's Readability Statistics, though in a limited way (DuBay, 2004). However, there are also some negative points to the formulas. Firstly, they are not entirely accurate when compared with human judgement (DuBay 2004). By only including superficial textual elements in a formula, it cannot calculate an entirely correct readability score for the text. The textual characteristics are important to calculate the readability, but characteristics of the reader need to be taken into account as well. These individual characteristics can be measured by conducting research into the reading process. An example is the reading time, which is different for people. Jakobsen & Jensen (2008) have already conducted research using eye tracking to gain insight into the cognitive effort for reading with different purposes. Secondly, there is no consensus on using one formula to predict readability. There are many different formulas that use different variables with different weights. This means that two different readability formulas can give a different readability score, which makes it difficult to accurately predict the readability of a text.

In the 21st century there have been some improvements in the new formulas that are being developed (Collins-Thompson, 2014). Modern readability formulas differ from the traditional formulas because they are no longer limited to superficial textual characteristics. Now they also include more complex lexical features, but also semantic features and discourse features, such as the degree of referential cohesion (Collins-Thompson, 2014). These improvements lead to more accurate scores, making the formulas more reliable to define the difficulty of a given text.

2.1.3. Indicators of reading difficulty

There is also no consensus yet about which elements predict reading difficulty. Similar to the readability formulas, there are many different classifications of readability indicators, though the indicators themselves are often similar. To test if these indicators have an influence on the readability of a certain text, specific features in a text will be measured. Readability features are an operationalisation of readability indicators.

In their research, De Clercq & Hoste (2016) make a classification of features of reading difficulty. To classify the features, they first identify the general indicators of readability and differentiate between four groups: vocabulary, structure, coherence and other. They group the defining features together in four different groups: traditional features, lexical features, syntactic features and semantic features.

The traditional features are lexical and syntactic features that have been used in other research and in the classical readability formulas, and that have been proven to be able to predict readability. When testing which features are the best readability predictors, De Clercq & Hoste found that there are differences between English and Dutch. For example, the average number of content and function words is a good predictor for English, but not for Dutch. It is not surprising that there are differences between two languages, since different languages use different grammatical structures and language patterns. However, there are also some similarities in good readability predictors for both languages, with Term Frequency-Inverse Document Frequency (tf-idf) being a good predictor of readability in both English and Dutch.

Kraf & Pander Maat (2009) use a different classification. They start by identifying the different indicators that can predict readability and identify six different categories of indicators: vocabulary difficulty, complexity of the sentence, information density, coherence, concreteness and personality. With personality they do not refer to the individual reader characteristics that also have an influence on readability, but this category refers to how personal a text is. For each indicator, they also identified several features that can be used for measuring the indicator. For example, to see if vocabulary difficulty has an influence on the readability, the word length or the word frequency can be calculated. Kraf & Pander Maat found that the feature "word length" has less influence on readability as the reader becomes older. This is remarkable because features related to word length are often used in readability formulas, such as the Flesch-Kincaid Readability formula and the SMOG Formula (Flesch, 1948; McLaughlin, 1969). De Clercq & Hoste (2016) also classified this feature with the traditional features, indicating that the feature had already been proven to be useful. So this feature is still used very often, and the decrease in the influence of word length with age does not mean that this feature is not at all useful anymore.

When comparing these two different classifications, there are differences but also some similarities. Both research teams have a similar classification for the readability indicators, with Kraf & Pander Maat (2009) being more elaborate. Indicators that are recurring are vocabulary, coherence and sentence structure. The fact that these indicators are recognised by more than one researcher shows that these are useful indicators for readability prediction. Furthermore, these are also the indicators that are mostly used in the readability formulas. The Flesch Reading Ease formula, for example, uses average sentence length and average number of syllables per word to predict readability. These features can be classified under sentence structure and vocabulary respectively.

2.2. Translatability

2.2.1. Definition

As was the case with readability, there are multiple definitions for translatability. Underwood & Jongejan (2001) define translatability as follows: “The notion of translatability is based on so-called “translatability indicators” where the occurrence of such an indicator in the text is considered to have a negative effect on the quality of machine translation” (p. 363). This definition is specifically focused on machine translation, but also applies to human translation. When a text is translated by a person, there are also certain indicators that have an impact on the quality of the output. Campbell (1999) proposes certain indicators of translatability in his research, which will be discussed later. Sun (2015), however, defines translatability as “the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria” (p. 31). Both definitions focus on different aspects of translation difficulty. The definition of Underwood & Jongejan states that translation difficulty is caused by certain indicators in the text, much like readability. Contrarily, Sun mentions that the translatability of a text is indicated by how much cognitive effort is needed to translate the text.

From these definitions it can be deduced that translatability is similar to readability because certain indicators in a text can give an indication of the translatability of a text. There is a direct connection between readability and translatability, because readability can be seen as a part of translatability: in order to translate a text, the translator first needs to understand the source text. The indicators of readability can therefore also be seen as indicators of translatability. However, translatability is more than simply calculating the readability. Translating a text consists of different phases. Vandepitte (2016) identifies six steps when translating a text, starting with the exploration of the translation event and ending with the revision and review of the target text. Translatability can be measured during the reading and translation phase. Readability, however, only provides information about the reading phase, but does not include information about the translation phase, which will have to be collected in a different way. The difference between readability and translatability is that readability is concerned with reading for comprehension, while translatability with reading for translation. Jakobsen & Jensen (2008) studied the differences in cognitive effort required for these two reading purposes. They found that the purpose of reading has an influence on the task time, the fixation count and gaze time. Fixation count is how many times the participant involved fixated on certain points in a text and gaze time is how long all these fixations lasted during the task at hand (Jakobsen & Jensen, 2008, p. 107). More specifically, reading with the intent of translating required more cognitive effort. Because translating involves two different languages, Jakobsen & Jensen believe that pre-translation starts when the translator starts reading the source text. The difference between readability and

translatability thus consists also of the fact that readability has to do with a monolingual process, but translatability is concerned with two languages (Vanroy, De Clercq & Macken 2019). Because these two concepts are rather similar, though clearly distinguishable from each other, the definitions also show some similarities, while also highlighting the differences.

2.2.2. Indicators of translation difficulty

Much like readability, certain textual characteristics in the source text can be seen as predictors of translatability. Campbell (1999) recognises the following translatability indicators: meaningfulness, complex noun phrases, abstractness, and frequency and familiarity. The meaningfulness category refers to word combinations, such as collocations, in which the translation of a word is dependent on the other words in that group, because on its own the word is low in meaningfulness. Underwood & Jongejan (2001) use another classification, sharing some similarities with Campbell's classification. In the classification of Underwood & Jongejan a distinction was made between general indicators and system-specific indicators. The general indicators are indicators that have already been studied in other research, including lexical ambiguity. The second category are indicators that are specific for the machine translation system they used in their study and includes features such as sentences over 25 words with at least one adverb.

One of the indicators of Underwood & Jongejan's classification is compounds comprising three or more nouns. This can be compared to the complex noun phrases category of Campbell (1999). Complex noun phrases are difficult to translate into Dutch because the grammatical structure is different in Dutch, making it more difficult to find a good translation. One indicator that is only present in the classification of Underwood & Jongejan is sentence length; it does not appear in Campbell's classification. The differences in the classifications indicate that there is no consensus yet on the best translatability indicators and that more research is still needed. Another explanation can be that the classifications are dependent on the purpose of the study at hand.

Some of the indicators that are used for predicting translatability have also been identified as important for predicting readability. Campbell (1999) suggests frequency as one translatability indicator, the hypothesis being that more frequently used words are easier to translate than less common words. Frequency is also found to be an effective predictor of readability by Kraf & Pander Maat (2009). Campbell, however, urges caution with this translatability indicator as it is dependent on the corpus that was used to compile the list and frequency lists are compiled for native speakers. Second language speakers, however, might be more familiar with other words and hence frequently used words of the latter are likely different than for native speakers. This indicator is also absent from Underwood & Jongejan's classification, showing that this indicator still needs more research in order

to know how effective it is to predict translatability. For readability this is not a problem, since readability is concerned with one language that is the mother tongue of the reader.

Abstractness is another indicator that is used for readability and that is also relevant for translatability. Campbell (1999) specifically focuses on abstract nouns as being more difficult to translate, whereas Kraf & Pander Maat (2009) focus more on the text as a whole, with concrete texts being easier to read than abstract ones. So for translatability the focus is more on abstractness within the text, while readability focuses more on abstractness on the level of the text and does not focus on smaller elements in a text. Thus the indicator is important for both readability and translatability, only with a slightly different focus. Similar to the readability indicator sentence complexity (Kraf & Pander Maat, 2009), Underwood & Jongejan (2001) also distinguish features concerning prepositional phrases and subordinate clauses and their position in the sentence. Sentence complexity has therefore an influence on both reading and translating. If a sentence is complex, it will be more difficult to understand what the sentence means and will thus also take more time to translate the sentence.

There are also differences between the indicators and features of readability and translatability. For example, for readability there is no complex noun phrase indicator. A feature that is, however, included in many readability classifications and formulas to quantify the vocabulary difficulty is average word length. This is not exactly the same as the translatability indicator. Complex noun phrases can consist of short words, thus not qualifying as being difficult for readability. However, they both show that vocabulary difficulty influences both translating and reading. Both classifications of translatability indicators mentioned here also do not include any indicators of coherence.

2.2.3. Research

Unlike readability, translatability has not yet been the subject of much research. There is also no consensus yet on how to measure translation difficulty. To assess the translatability of a text, researchers focus either on the translation product, much like readability, or on the translation process, using eye tracking and keyboard logging to collect data, or on both the product and the process.

2.2.3.1. Translation product

There are two ways in which the translation product can be used to assess how difficult a text is to translate: by examining the translation variation or the translation quality.

The first way of assessing translatability is by looking at the translation variation. With translation variation is meant how many possible translations there are for a source token. This variation in translations can give an indication of how difficult a text is to translate. If a token has many different possible translations, then more mental effort is required to make a decision of which translation is

best. To examine translation variation, Campbell (2000) introduced Choice Network Analysis (CNA). CNA is a method of mapping the mental process while translating, represented by the choices that translators make when translating a text. When different people translate the same text, these different translations can be combined into a network, showing all the different choices between which they had to choose. The more elaborate the network is, the more possible translations there are and thus the more difficult a source token is to translate. CNA is therefore a useful method of estimating the translation difficulty of a given text. By mapping all the different choices that were made for source tokens the translatability of a text as a whole can be deduced and it also shows which types of source tokens are more difficult. Word translation entropy is also linked with CNA. Word translation entropy is a way to quantify how many choices a translator has when translating a source token. If there are many possible alternative translations this will result in a higher word translation entropy score. CNA can be seen as a way to visualise word translation entropy.

Another way of using the translation product to assess the translatability of a text is by assessing the quality of the translation or how many errors a translation contains. This method, though often used, is also criticized for being not objective, since the assessment is dependent on the assessor. One or several assessors will correct a translation and highlight the mistakes, but this correction may vary from assessor to assessor.

In their research, Eyckmans & Anckaert (2017) compared two different assessment methods, namely the Calibration of Dichotomous Items (CDI) method and the Preselected Items Evaluation (PIE) method, in order to find out which of the two methods was the most objective and reliable one. The difference between the CDI and the PIE method is that for the PIE method, before the text is translated, the assessors are first expected to make a list of 10 items that they think will be difficult to translate for translator trainees. A list is then made of these items with correct and incorrect translations. For the assessment of the translations, the assessors focus only on the items on this list. In the CDI method this list is not made in advance. This method includes a pre-test phase in which a sample of translator trainees translate the text and then the assessors look which are the difficult items based on the translations. In the CDI method the assessors thus base their list of difficult items on the evaluation of actual translations. Eyckmans & Anckaert concluded that the CDI method was the more objective and reliable method of the two. The experiment showed that with both methods there was little agreement between the assessors, who selected many different difficult items, but with the CDI method there was some more overlap than with the PIE method. When using the PIE method, the assessment is more dependent on the assessors than when using the CDI method, because the assessors not only evaluate the translation, but they are also the ones who elect the items that will be focused on for the actual assessment of the translation. In the CDI method this list is based on actual

errors made in translations of a sample of trainees. Furthermore, using the PIE method may result in missing more errors in the translations. Because this assessment method focuses on a predefined list of difficult items that were selected before translating, other errors that were made that are not resulting from one of the items on the list will not be taken into account. The results from this experiment show that in order to make an objective judgement about the translation quality multiple assessors are needed.

2.2.3.2. Translation process

Another way of assessing translation difficulty is by analysing data about the translation process. Data about the translation process represent the cognitive effort that is needed to translate a text. The more cognitive effort that is required to translate a text, the more difficult that text is. Common methods of retrieving information about the translation process are eye tracking, keyboard logging, screen recording and think-aloud protocols (Sun, 2015).

Vanroy, De Clercq & Macken (2019) conducted research trying to see if there is a correlation between product and process data. They focused on process features that are related to duration, revision and gaze. As product features, they used the error count, word translation entropy and syntactic equivalence. These product data serve as a proxy for translation difficulty. They concluded that there is a correlation, showing the validity of process data for measuring translatability. More specifically, they found that the features average pause ratio and the number of revisions are strongly correlated with the product data. When a part of a text is more difficult, the translator might take longer pauses to think about the translation and make more adjustments afterwards. The other features they studied also showed a correlation, though less significant. This study shows that both product and process data can be used to measure translatability, because both data lead to similar results.

Liu, Zheng & Zhou (2019) examined what impact the complexity of the source text has on the cognitive effort required to translate a text. The process features they used were all collected using eye tracking, focussing on two specific features, namely fixation and saccadic duration (FSD), and pupil dilation. A saccade is a quick movement of the eyes when looking for a new part of a text to focus on. The product features in their study were readability, word frequency and non-literalness. The participants in this study were 26 master translation students who were asked to translate three English tasks into Chinese. They also had to assess the texts before and after the translation. The texts used in this experiment are part of the multiLing dataset, a dataset based on six English texts that are translated into different languages, while logging the keyboard activity and sometimes also tracking the eye-movement. In accordance with Vanroy, De Clercq & Macken (2019), they could conclude that there is a correlation between product and process data. However, pupil dilation was deemed not reliable to measure translatability, because it seemed to be less influenced by the difficulty of the source text

and more susceptible to other factors, such as the order in which the texts were presented. Liu, Zheng & Zhou also tested if the estimated readability and the process data correlated with the subjective rating of translatability. Both before and after the translation of the texts the students were asked to rate the translatability of the texts. The students involved in the experiment did rate the translation difficulty of the texts in accordance with both the readability measures and the process data. Therefore subjective rating of the translatability can be used to estimate the readability of said text and to give a general indication of the translatability of a text as a whole. Process data can then show in more detail which specific elements in the text are more difficult.

The results of Liu, Zhen & Zhou (2019) are similar to those of Vanroy, De Clercq & Macken (2019). The studies examined what the correlation is between product and process data. Whereas the features used as process and product data differed between the two studies, the conclusion of the two experiments showed that process data and product data are both reliable for measuring translation difficulty. This shows that both methods can be used to measure translatability. Moreover, the experiment of Liu, Zheng & Zhou demonstrated that translators themselves can also accurately rate the translatability of a text, making this also a viable option for measuring translatability.

2.3. CRITT Translation Process Database

2.3.1. What?

The Center for Reasearch and Innovation in Translation and Translation Technology developed a database that contains data from many different studies about text processing, such as translation, post-editing and copying, called the CRITT Translation Process Database (CRITT TPR-DB) (CRITT, s.d.). The database is publicly available under a creative commons license. To collect this data, Translog, Translog-II or CASMACAT workbench is used. Translog is a program that allows you to record the keyboard activity of the participant. The program can also be used in combination with an eye tracker, which will then also allow you to track the eye movement of the participant. Both keyboard logging and eye tracking are useful to gather information about the cognitive effort required to translate a text, as shown in section 2.2.3.2. Currently the database contains data from more than 3000 sessions. The database also offers the option to download post-processed versions of the raw data in the form of multiple tables that are easier to process.

One of the datasets in this database is the multiLing dataset (CRITT, s.d.). This dataset is based on six English texts, four of which are news articles and two are sociological texts from an encyclopaedia (CRITT, s.d.). The studies that contributed to this dataset performed various different tasks, such as from-scratch translation, post-editing and copying. Currently the dataset contains data from 26 studies, involving eight different languages, such as Japanese, Spanish and Arabic.

2.3.2. Research

Since the beginning of the 21st century researchers have been contributing to the expansion of the CRITT Translation Process Database. Below we summarise some of the important contributions.

One contribution to the multiLing dataset was made by Hvelplund (2011). He conducted research into the cognitive resources that are used during translation, using keyboard logging and eye tracking to collect data. He specifically examined how cognitive resources are distributed, how they are managed and how the cognitive load changes during the translation process. The three features he focused on were total attention duration, attention unit duration and pupil size. For his experiment, Hvelplund asked 12 professional translators and 12 student translators to translate three English texts with a varying degree of complexity into Danish. The three texts in this study are text 1, Killer nurse receives four life sentences, text 2, Families hit with increase in cost of living, and text 3, Spielberg shows Beijing red card over Darfur, of the multiLing dataset. Two of the three texts had to be translated under a time constraint. Both eye tracking and keyboard logging were used to collect data about the translation process. Text 1 and 3 are also used in the present research

One of the conclusions of his research is that translating a difficult text does not influence the cognitive load required. This conclusion is in contradiction to what Vanroy, De Clercq & Macken (2019) and Liu, Zhen & Zhou (2019) concluded in their respective studies. However, Hvelplund (2011) also mentions that this outcome could be explained by the features that were used in his research, more specifically pupil size, since this is considered not to be a reliable feature to measure cognitive load. It is a feature that is more often used to measure how surprising something is for a participant, rather than how difficult something is. This remark is in line with the findings of Liu, Zheng & Zhou (2019), who discussed that pupil size is not a reliable feature to measure the translatability of a text, since it was also dependent on other features, such as the order in which the texts were presented. Another remarkable result was that there was no statistical difference in the time spent translating an easy or a difficult text. However, Vanroy, De Clercq & Macken (2019) could conclude that the difficulty of a text, as represented by product features, did change the duration of translation.

Bangalore et al. (2015) studied the effect of syntactic variation on the behaviour of the translator during translation and post-editing. Syntactic variation or entropy can give an indication of the translatability of a text, just like word entropy, but on a different level. Entropy is used to quantify how similarly a text is translated by different translators. Differences can occur on syntactical level, looking at the syntactical structure, and on word level, looking at the lexical choices that were made. The word entropy of a text shows which specific tokens are more difficult to translate, while syntactic entropy will reveal which sentences are more difficult. For his study Bangalore et al. used different datasets from the CRITT Translation Process database, involving three different languages of

translation, namely Danish, German and Spanish. The process features they focused on were the total production time without pauses and the coherent typing activity. Vanroy, De Clercq & Macken (2019) use a similar production duration feature in their research, suggesting that this feature is likely a good indicator of translatability. Bangalore et al. found that syntactical entropy, used to represent syntactic variation, had an effect on both features that were used for the translation to all the target languages in this study. This effect could not be found for post-editing.

The study of Bangalore et al. (2015) showed how a more difficult text has an influence on the translation duration. However, Hvelplund (2011) concluded that a more difficult text did not require more time to translate. The difference in results could possibly be explained by the different features that were used in both studies. Also the fact that Hvelplund used pupil dilation as one of the features could have influenced the results, leading to a conclusion opposite of Bangalore et al. and Vanroy, De Clercq & Macken.

3. Research Objective

The general context of this master thesis is the following: what makes a text difficult to translate and what is the relationship between readability and translatability? More specifically, we investigate the following questions:

1. Is the readability of a text a good indicator of the subjective classification of translatability? More concretely, if we have two texts where one is classified more readable than the other, do people also mark the more readable text as more translatable, and vice versa?
2. What are the typical elements that make translation difficult? If we ask students and professors to mark the difficult elements in the texts (sentences, word groups, individual words), are there items in common between different participants and between teachers and students, and, if so, can we link these common elements to the translatability indicators that were found in the related research described in section 2.2.2?
3. Can we deduce translation difficulty from objective process data? Specifically, if we collect process data during the translation process, can we estimate the translation difficulty based on some process features and can we link these to specific textual elements that were marked as being difficult in point 2 above? The process features that we focus on here are translation duration and number of textual revisions.

4. Methodology

In this chapter we explain the methodology that was used in this experiment in order to find an answer to the research questions stated in the previous section. First we briefly summarize the process of the data collection: what are the steps of the data collection, what is their interrelationship and how do they relate to the research objectives? Then we list the texts that were used in this process. Next, we describe which participants took part in the experiment. Finally, we elaborate on the features that were the focus of our work: the subjective ratings, the difficult items and the process data. We also discuss how we normalised the data in order to remove participant-specific bias.

4.1. Data collection

For the collection of the data, two English texts from the multiLing dataset were selected, one easier text and one more difficult text according to readability scores and previous research. The texts were then evaluated by translation teachers and by students, using questionnaires. The teachers were also asked to indicate specific items that they thought would be difficult for students. The students were first asked to fill in the Lexical Test for Advanced Learners of English (LexTALE) (http://www.lextale.com/), which is available online. This test is intended to measure the level of English knowledge of non-native English speakers and consists of a lexical decision task. Next, the students were first asked to copy a text in Translog and then to translate the two selected texts, while the translation process was being logged with Translog. During the translation the students were not allowed to use any resources. Afterwards they also had to indicate what they would normally look up and give a short commentary about these items.

4.2. Texts

The two texts that were used in this project have been taken from the multiLing dataset (see appendix A and B). The dataset contains six texts. The rationale for the selection of these two texts was to select one easier and one more difficult text. The two texts that were selected are text 1, Killer Nurse Receives Four Life Sentences, and text 3, Spielberg Shows Beijing Red Card over Darfur. These two texts are the easiest and most difficult text in the dataset respectively. To estimate which texts were the easiest and the hardest we used the readability demo of the University of Ghent. Additionally, the average time (normalised) needed for translation of the text into other languages, such as Spanish and Japanese, was also calculated based on the data from the multiLing dataset.

The readability demo of the University of Ghent (https://www.lt3.ugent.be/readability-demo/) is available online and provides information about different aspects of a text that can be linked to the readability of a text, such as the average word length, the type token ratio and the average number of subordinating conjunctions. Furthermore, the demo also calculates the readability score of a text of seven different formulas, amongst which are the Flesch Reading Ease formula and the Gunning Fog

Index. For the selection of the two texts, we focused on the readability scores of the different formulas. For six of the seven formulas text 1 received the lowest score, which means that text 1 is supposedly the easiest text according to these formulas. For one of the seven formulas, the Flesch Reading Ease formula, text 1 received the highest score. However, the Flesch Reading Ease formula has a different scale, which means that for this formula the higher the score, the easier the text is. Therefore, all seven formulas indicated that text 1 is the easiest text. Text 3 received the highest score for four formulas. This means that, according to these formulas, text 3 is the most difficult text. No other text received the worst score for more than four readability formulas. Therefore, text 3 was selected as the most difficult of the six texts.

	T01	T03
Flesch Reading Ease formula	73.7	35.98
Flesch-Kincaid Grade Level	6.41	15.1
Dale-Chall Reading Grade Score	8.79	11.99
Coleman-Liau index	8.87	14.16
Gunning Fog Index	8.95	18.29
SMOG	9.34	15.9
ARI	6.92	17.18

Table 1 - Readability scores T01 and T03

The translation duration per token was calculated using the data available in the multiLing dataset. Of nine studies that include translating the texts, the data about the translation process was available to download. For seven of these studies all of the six texts from the dataset were translated. From these seven studies, the average normalised translation duration on session level was compared. The session duration starts when the session in Translog is started and ends when the program stops logging the keyboard activity. In most of the studies, text 1 turned out to take the least time to translate and text 3 and 4 both seemed to take the most time. In this experiment we selected text 3, because this text was also considered to have the lowest readability.

Studies	Language pairs	T01	T03
BML12	English - Spanish	3212.66	6072.94
ENJA15	English - Japanese	6450.78	11049.27
NJ12	English - Hindi	6864.2	12841.06
RUC17	English - Chinese	5455.13	8163.78
SG12	English - German	6145.33	8890.99
STC-17	English - Chinese	5384.5	8095.46
WARDHA13	English - Hindi	11034.69	19528.19

Table 2 - Translation duration per token of the multiLing studies (in milliseconds)

The order in which the two texts for translation were presented to the students varied: for half of the students T01 was the first text and for the other half T03 was the first text. This difference in

presentation order was done to avoid that the results of the experiment were influenced by the order in which the texts were presented to the participants.

Before translating the two selected texts, the students were asked to copy a text from the multiLing dataset in Translog-II. The purpose of this copy task was to make the students familiar with the program. For the copy task, text 4, Climate Change (see appendix C), was selected. Since this text was not used for translation, it was not selected based on its readability score or the translation duration into other languages. Instead text 4 is the shortest text out of the six texts, which is sufficient to allow the participants the time to get used to the program, while not requiring too much time.

4.3. Participants

The participants in this experiment can be divided into two group: teachers of English translation and translation students. In total, nine teachers participated in the study. The teachers, selected from a list of universities that have a translation program, were contacted via e-mail. In total 25 teachers were contacted, of which nine contributed to the study. The teachers work at varying universities in Belgium, the Netherlands and the United Kingdom, but all of the teachers speak Dutch. Their teaching experience ranges from 5 to more than 20 years.

The student participants are students who are following the Master in Translation at Ghent University or who are following the postgraduate course Computer-Assisted Language Mediation (CALM) at Ghent University. All the students in the master program who are studying English and all the students in the CALM postgraduate course who had previously studied English in the Master in Translation were contacted via e-mail. In total, ten students participated in the study, four students finishing their master and six students following the CALM postgraduate course. The students following the postgraduate course had previously finished the Master in Translation at Ghent University. For all students, English is or was one of their elected languages.

The participants were each assigned a letter and a number. The letter ‘D’ is used for the teachers, in combination with a number from 01 to 09. For the students the letter ‘P’ is used with a number from 01 to 10.

4.4. Data

In the experiment we collected three different types of data: subjective ratings, difficult items and translation process data. In the following sections we discuss what each type of data consists of.

4.4.1. Subjective rating

The first type of data that was collected is the subjective rating. The teachers and the students were both asked to fill in a questionnaire with general questions and with questions specifically about the two selected texts. The two groups of participants received slightly different questionnaires, which

are included in appendix D and appendix E. The questionnaire for the teachers consists of 11 questions, of which five asked about more general information, such as at which institutions they teach courses, and six questions about the translatability of the texts, such as why they considered one text more difficult to translate than the other text. The questionnaire for the students contains 13 questions, nine more general questions, such as which languages they were studying, and four questions about the translatability of the texts, such as which kind of items were most difficult in the texts. Both questionnaires contain the question to rate the difficulty of text 1 and text 3 from 1 to 10, where 1 means that the text is easy and 10 that the text is difficult. The teachers also had an additional question to indicate if they would let students translate the texts in the first, second, third bachelor, or master year.

4.4.2. Difficult items

The second type of data consists of items marked as difficult by the participants. The teachers and students both indicated difficult items in the two texts. More specifically, the teachers were asked to indicate items that they thought that students would have difficulties with when translating. They had to indicate these items without translating the texts. The students, however, first translated a text and when that text was translated, they were asked to indicate items that they would normally have looked up, because they were not allowed to use any resources while translating. These items can also be considered difficult items.

We compared the difficult items that were indicated and their accompanying comments to see if the teachers and students recognised the same difficulties. This is by nature a subjective activity, because, for instance, they did not always indicate exactly the same words, but rather an overlapping range of words. For those difficult items, we looked whether the core of the difficult items was the same. Still, most of the time it was rather clear what items were in common for the different participants. The difficult items of both texts can be found in appendix F and G.

4.4.3. Process data

The tables containing the data that was recorded during the translation process by Translog consist of a lot of information. There are, for example, tables with data about the text as a whole, with keystroke data, with data on segment level, and so on. In this master thesis, we will be focussing on the segment tables (.sg) and session tables (.ss). Based on the research discussed in section 2.2.3, we decided to focus on the following features:

- Dur**:** Dur stands for duration and in this research we will look at this feature in both the session and the segment tables. In these tables, the feature has a slightly different definition. When looking at the entire translation process, Dur refers to “the production duration of a final target text per session” (CRITT, s.d.). The duration starts when the session in Translog is

started and ends when the session is stopped. The duration on segment level, however, shows how long it took to translate one segment and "counts the time from the first keystroke on a particular segment to the last keystroke relating to that particular segment" (CRITT, s.d.). If the segment was revised several times then the times are added together. This can be considered a good measure of how much time it took for the student to translate a given segment. Presumably this is a meaningful measure of the translation difficulty of the segment.

- Nedit**:** This feature can be found in the segment tables and is defined as the "number of times the segment was edited" (CRITT, s.d.). This feature can be classified as a revision feature. The more times a token or segment is edited, the more difficult it can be assumed to be. Important to know is that the default value for Nedit 1 is. This means that if a student has an Nedit value of 1, the student translated the segment, but did not come back to the segment to edit it later. If a student has an Nedit value higher than 1, then the student did return to the segment to change or add something to the translation.

4.4.3.1. Duration (Dur)

There is a considerable amount of variation in translation speed between the different students. The total translation duration of both text 1 and 3 combined is 721,296 milliseconds for the slowest student and 2,809,437 milliseconds for the fastest student. Therefore the data needed to be normalised in order to have a representation of the relative translation duration per segment in order to compare different students.

The total duration on session level for a given student, which is the overall duration for text 1 and 3 combined, was divided by the number of tokens in the texts. The total token count is 306 tokens, 160 in text 1 and 146 in text 3. The number of tokens in the texts is higher than the word count, because tokens also include punctuation marks. This gives the average duration per token for a participant (DurT) (see Table 3).

In Table 4 we calculated the normalised translation duration for the two texts together by adding up all the durations from the segment level. What can be noticed is that this duration is not the same as when the normalised duration was calculated starting from the duration on session level. An explanation is that the duration on segment level only starts when the student actually starts typing the translation of a segment and stops when the typing is stopped, while the duration on session level starts when the session in the program is started and ends when the session is stopped. Therefore, the duration on session level also includes pauses that the students might take in between translating segments. Even though the two normalised durations are different, the difference is not that big: P01 has the biggest difference with the normalised duration started from session level being 9.78% longer than the duration started from segment level, which equals a duration that is 343.50 milliseconds

longer. The other participants have even smaller differences between the two normalised durations. Because the difference is that small, we decided to only use the normalised duration started from session level in the data analysis.

Participant	Duration	DurT
P01	1074860	3512.61
P02	1802204	5889.56
P03	1663547	5436.43
P04	2809437	9181.17
P05	944733	3087.36
P06	2179672	7123.11
P07	2334203	7628.11
P08	721296	2357.18
P09	1377421	4501.38
P10	1460782	4773.8

Table 3 - Average duration per token – session level (in milliseconds)

Participant	Duration	DurT
P01	969750	3169.12
P02	1659219	5422.28
P03	1608109	5255.26
P04	2625796	8581.03
P05	917562	2998.57
P06	2103516	6874.24
P07	2236797	7309.79
P08	708031	2313.83
P09	1291766	4221.46
P10	1440719	4708.23

Table 4 - Average duration per token - segment level (in milliseconds)

Next, the duration for each segment from the segment table was divided by the number of tokens in that segment. This gives the average duration per token for that particular segment per participant (DurS). Finally, DurS was divided by DurT and that gave the normalised duration per word for a particular segment for each student, which can then be used as a possible indication of the translation difficulty of the segment. It should be noted that this normalised duration is not expressed in milliseconds or seconds, but it is a ratio of segment time divided by total time, hence these values are dimensionless.

This normalised duration gives a profile of a student and varies per student. We can visualise the variation in the normalised translation duration in a box-whisker plot (see Figure 1, in section 5.3.1.1), which shows the following information:

- The “x” inside the boxes represents the average value of the normalised duration per token for that specific student.
- The middle horizontal line in the box is the median value. This line divides the data into a bottom half and a top half.
- The bottom line of each box represents the median of the bottom half, which is the first quartile value (Q1). The top line of each box is the median of the top half, which in its turn is the third quartile value (Q3). The difference between these two lines is the inter-quartile range (IQR). It corresponds to the height of the box.

- A data point is classified as being an outlier if it exceeds a distance of 1.5 times the IQR below the first quartile Q1 or 1.5 times the IQR above the third quartile Q3. If there are one or more outliers then the bottom or top whisker under or above the box respectively shows the value of the 1.5-limit, while the outliers are shown as dots. If there are no outliers then the bottom or top whisker shows the minimum or maximum value in the dataset.

In order to be able to quantify the data analysis each value was labelled with a number in comparison to the quartile values of the normalised translation duration of each student. The numbers used to classify the data points range from 1 to 6. Table 5 below shows the description of each class.

Class	Description	
1	much faster than the median	low outlier
2	faster than the median	lower than Q1, but not outlier
3	a bit faster than the median	between Q1 and Q2
4	a bit slower than the median	between Q2 and Q3
5	slower than the median	higher than Q3, but not outlier
6	much slower than the median	high outlier

Table 5 - Classification data points

4.4.3.2. Number of edits (Nedits)

The other process feature that was focused on in this study is the number of edits (Nedits). This feature was not normalised by dividing the values by the number of words. This is because there is no clear connection between how many edits are carried out and the length of the segment. It is possible that a very long sentence is easy to translate and vice versa.

5. Results

This section discusses the results of the different phases of the experiment: the overall subjective rating of the two texts, the labelling of difficult items and the analysis of the process data.

5.1. Subjective rating

Both the teachers and the students were asked to give their personal opinion about the translation difficulty of the two texts. Specifically, one of the questions of the questionnaire was to rate the difficulty of text 1 and 3 from 1 to 10, 1 being easy and 10 being difficult. Table 6 shows the subjective ratings of the teachers and Table 7 shows the ratings of the students.

	T01	T03
D01	5	7
D02	2	4
D03	1	3
D04	6	7
D05	4	8
D06	5	4
D07	2	3
D08	6	8
D09	7	8
Average	**4.2**	**5.8**

Table 6 - Subjective rating teachers

	T01	T03
P01	3	4
P02	3	8
P03	2	6
P04	5	7
P05	4	8
P06	4	6
P07	4	8
P08	3	8
P09	3	7
P10	4	7
Average	**3.5**	**6.9**

Table 7 - Subjective rating students

All the teachers, except D06, indicated text 3 as being more difficult than text 1. The average score given to text 1 is 4.2 and the average for text 3 is 5.8. On average, the difference between the two ratings is 1.56. The reasons for rating text 3 more difficult were vast, with recurring comments being that text 3 is denser, has more complex structures and requires more background information.

The students were also asked to rate the difficulty of the two texts, but only after they had translated both texts. All the students agreed that text 3 was more difficult to translate than text 1. The difference between the two texts is 3.4, with an average score of 3.5 for text 1 and 6.9 for text 3. The students also gave many different reasons for finding text 3 more difficult. Some recurring comments refer to the more complex structure, the topic of the text, namely the Darfur conflict in 2008, the specific terminology and the cultural references.

	T01	T03
D01	1	2
D02	1	2
D03	1	2
D04	2	3
D05	2	3
D06	2	2
D07	2	2
D08	1	2
D09	2	2
Average	**1.56**	**2.22**

Table 8 – Year of study chosen by teachers

An additional question given to the teachers was in which year they would have the texts translated by their students (see Table 8). All the teachers indicated that they would let students in the bachelor years translate the texts. No teacher chose for the master. For text 1, four teachers would let the text be translated in the first bachelor and five teachers would give the text in the second bachelor. The year in which they would have text 3 be a task varied between the second bachelor and the third bachelor, with most teachers choosing the second year. Additionally, three teachers would let both texts be translated in the same year, though they rated the difficulty of the two texts differently.

5.2. Difficult items

Both the teachers and the students were asked to indicate items that are difficult to translate. The teachers had to indicate items that they thought would be difficult for students and the students indicated the items that they found difficult while translating. In the next section, we look at the segments that are considered most difficult based when looking at the number of difficult items per segment. After that we discuss the difficult items on item level.

5.2.1. Segment level

5.2.1.1. Text 1 – Killer Nurse Receives Four Life Sentences

T01	number of teachers	total number of difficult items	percentage of all difficult items
seg01	7	12	16.22%
seg02	4	7	9.46%
seg03	3	5	6.76%
seg04	9	12	16.22%
seg05	2	3	4.05%
seg06	3	3	4.05%
seg07	5	8	10.81%
seg08	6	9	12.16%
seg09	5	9	12.16%
seg10	3	3	4.05%
seg11	3	3	4.05%

Table 9 - Difficult items T01 - teachers

T01	number of students	total number of difficult items	percentage of all difficult items
seg01	3	3	7.89%
seg02	7	7	18.42%
seg03	4	4	10.53%
seg04	6	9	23.68%
seg05	4	4	10.53%
seg06	1	1	2.63%
seg07	4	4	10.53%
seg08	2	2	5.26%
seg09	1	1	2.63%
seg10	2	2	5.26%
seg11	1	1	2.63%

Table 10 - Difficult items T01 - students

Table 9 above shows how many teachers recognised difficult items in which segments of T01. It also indicates how many difficult items were highlighted in total per segment and what percentage it represents of all the difficult items in text 1. Table 10 contains the same information, but for the students. The percentages in both tables were calculated for the teachers and students separately by dividing the number of difficult items per segment by the total number of difficult items that were indicated by either the teachers or the students in one of the two texts.

When looking at Table 9 there are two segments in text 1 with the most difficult items of teachers, namely segment 1, which is the title, and segment 4. Segment 1 contains 12 difficult items indicated by seven different teachers. Segment 4, however, contains the same number of difficult items, but each teacher marked at least one difficult item. In both segments the number of difficult items represents 16.22% of all the difficult items in text 1. Table 10 shows that the segment with the most difficult items from students is segment 4, which the teachers also considered as one of the most difficult segments. In this segment, six different students indicated in total nine difficult items, amounting to 23.68% of all difficult items. In T01 there is also no segment in which all students highlighted a difficult item, though in segment 4 every teacher did recognise a difficult item.

In segment 4, all of the teachers indicated difficult items. This is the only segment in which every teacher indicated at least one difficult item. In segment 2 the most different students marked difficult items, with seven different students indicating difficult items. For the students there is no segment in which every student indicated at least one difficult item.

The segments with the least number of difficult items for the teachers are segment 5, segment 6, segment 10 and segment 11. These four segments contain three difficult items or 4.05% of all difficult items indicated by the teachers. For the students, the least difficult segments were segment 6, segment 9 and segment 11. In these segments only one student highlighted one difficult item, which is 2.63% of the total number of difficult items for students in T01. Segment 6 and 11 are considered least difficult by both the students and the teachers. The other easy segments, segment 5 and 10 for the teachers and segment 9 for the students, do not receive the same rating from the other group.

5.2.1.2. Text 3 – Spielberg Shows Beijing Red Card over Darfur

T03	number of teachers	total number of difficult items	percentage of all difficult items
seg01	7	12	14.29%
seg02	9	14	16.67%
seg03	9	27	32.14%
seg04	6	15	17.86%
seg05	8	16	19.05%

Table 11 - Difficult items T03 - teachers

T03	number of students	total number of difficult items	percentage of all difficult items
seg01	6	7	10.61%
seg02	8	13	19.70%
seg03	9	15	22.73%
seg04	8	15	22.73%
seg05	9	16	24.24%

Table 12 - Difficult items T03 - students

Table 11 shows the difficult items per segment of T03 for the teachers and Table 12 shows the same information for the students.

For the teachers, the most difficult segment is segment 3. This segment has 27 difficult items or 32.14% of all the difficult items in text 3. All the teachers highlighted at least one difficult item in this segment. The segment that contains the most difficult items according to the students is segment 5 with 16 difficult items from 9 students. This represents 24.24% of all the difficult items for students.

The segments in which all of the teachers indicated difficult items are segment 2 and segment 3. These segments contain difficult items from all 9 teachers, with in total 14 difficult items in segment 2 and 27 difficult items in segment 3. The most students, namely 9, recognised difficult items in segment 3 and segment 5. In text 3 there is no segment in which all the students indicated difficult items. However, every teacher did indicate at least one difficult item in segment 2 and 3.

The segment with the least number of difficult items is segment 1 for the teachers and for the students. More specifically, seven teachers indicated in total 12 difficult items, which amounts to 14.29%, while six students marked seven difficult items, which is 10.61%.

5.2.2. Item level

In this section we look at the specific difficult items. Some teachers and students also indicated a few times an entire segment as being difficult, because it is, for example, a long segment. However, in this section we look at smaller items, which is why we chose not to include the longer difficult items about an entire segment or subordinate clause in this section.

The teachers were asked to highlight items in both texts that they thought would be difficult for students to translate and they also gave a short explanation as to why a particular item was deemed difficult to translate. Table 13 shows how many difficult items each teacher indicated in the two texts and the total and average number of difficult items per text.. The students had to indicate items that they would have normally looked up, also including a short comment explaining why they would look up something. The number of difficult items per student, the total and average number of difficult items are included in **Error! Reference source not found.**.

Teachers	T01	T03
D01	6	5
D02	3	6
D03	20	13
D04	4	3
D05	3	11
D06	9	9
D07	5	5
D08	10	18
D09	9	14
Total	**69**	**84**
Average	**7.7**	**9.3**

Table 13 - Difficult items teachers

Students	T01	T03
P01	5	7
P02	2	8
P03	5	12
P04	2	3
P05	4	5
P06	5	10
P07	3	6
P08	2	5
P09	4	6
P10	4	6
Total	**36**	**68**
Average	**3.6**	**6.8**

Table 14 - Difficult items students

Looking at item level, the 9 teachers highlighted 69 difficult items in text 1 and 84 items in text 3. For the students the number of difficult items in text 1 is 36 and in text 3 this is 68. The averages for T01 are 7.7 for the teachers and 3.6 for the students. In T03, the teachers indicated on average 9.3 difficult items and the students 6.8. In both texts there is also a rather big variation in how many difficult items each teacher indicated. D03, for example, indicated 20 difficult items in T01, while D02 and D05 only marked three difficult items. The variation is not that big for the students. However, in T03 this variation is bigger than in T01. In T03, P04 only indicated three difficult items, while P03 highlighted 12 difficult items. In T01 the lowest number of difficult items is 2 and the highest is only 5.

	Teachers	Students
T01	2.09	1.06
T03	2.33	1.89

Table 15 - Average number of times difficult items were indicated

Table 15 shows the average number of times a difficult item was indicated by either the teachers or the students. On average a difficult item was marked by 2.09 teachers and by 1.06 students in T01. In T03 a difficult item was indicated on average by 2.33 teachers and 1.89 students. This average is rather low, because there are 9 teachers and 10 students who participated in the study, which means that the participants marked different difficult items. The average of the teachers is a little higher than the average of the students. Both for the students and for the teachers the average is higher for T03 than for T01.

Segment	T01	Teachers	Students	Total	Difference
1	**Killer Nurse**	6	3	9	3
1	**four life sentences**	4	0	4	4
2	**Hospital nurse**	2	5	7	-3
4	**four counts of murder**	8	6	14	2
7	**had been acting**	4	0	4	4
8	**the awareness of other hospital staff**	5	2	7	3
8	**put a stop to him and to the killings**	4	0	4	4
9	**have learned**	5	0	5	5

Table 16 – Selection of difficult items - T01

Table 16 shows a selection of the difficult items that were indicated in text 1. A full list of all the difficult items can be found in appendix F.

The items that are shown in this table are either the most frequent items or the items for which the difference between teachers and students is the biggest. Frequently highlighted difficult items are items that were considered difficult by at least half of the teachers or half of the students, meaning by five teachers or students. An absolute difference between teachers and students of 3 or more is considered a big difference, because it represents a third of the participating teachers or students.

For text 1 there are two instances where at least half of the students indicated the same difficult item, namely ‘hospital nurse’ and ‘four counts of murder’. This last word group is the item that was recognised as being difficult by most of the students, highlighted six times. For the teachers, there are four different difficult items that were marked by five teachers or more, namely ‘killer nurse’, ‘four counts of murder’, ‘the awareness of other hospital staff’ and ‘have learned’. The most highlighted item for the teachers is ‘four counts of murder’ and lines up with the most highlighted item of the students. The biggest difference between teachers and students is the item ‘have learned’. This item

was recognised as difficult by five teachers, but not by any of the students, amounting to a difference of 5.

Segment	T03	Teachers	Students	Total	Difference
1	**Darfur**	2	6	8	-4
2	**In a gesture sure to**	4	1	5	3
2	**rattle**	5	4	9	1
2	**Government**	4	0	4	4
2	**Beijing Olympics**	0	3	3	-3
3	**in the wake of**	3	0	3	3
3	**fighting flaring up again**	5	5	10	0
3	**set to**	3	0	3	3
3	**sought to**	3	0	3	3
3	**having close ties to**	5	2	7	3
4	**extensive investments**	3	0	3	3
4	**which includes one minister**	3	0	3	3
4	**crimes against humanity**	0	5	5	-5
4	**International Criminal Court in The Hague**	3	8	11	-5
5	**although emphasizing**	5	0	5	5
5	**Khartoum**	2	7	9	-5

Table 17 – Selection of difficult items - T03

Table 17 above shows a selection of the difficult items that were indicated in text 3, with the full list of difficult items included in appendix G.

In text 3, four difficult items were indicated by half of the teachers, namely 'rattle', 'fighting flaring up again', 'having close ties to' and 'although emphasizing'. These are the items that were most frequently indicated by the teachers, since there are no difficult items that were marked by more than five teachers. For the students, there are five difficult items that were indicated by at least half of the students, ranging from five times to a maximum of eight times, namely 'Darfur', 'fighting flaring up again', 'crimes against humanity', 'International Criminal Court in the Hague' and 'Khartoum'. The most frequently highlighted difficult item for them is 'International Criminal Court in The Hague'. The biggest difference between teachers and students can be found with the difficult items 'crimes against humanity', 'International Criminal Court in The Hague', 'although emphasizing' and 'Khartoum'. Three of these difficult items were more frequently highlighted by students and one of these items, namely 'although emphasizing' was more frequently highlighted by teachers, with students not marking this item as difficult.

We also tried to classify all the difficult items according to the classifications of Campbell (1999) and Underwood & Jongejan (2001), as discussed in section 2.2.2. However, many of the difficult items seemed to not fit in one of the categories. The teachers and students seemed to focus more on smaller

elements, such as words and their translations, and difficult noun phrases or compounds. In the literature discussed in section 2.2.2., however, there is more focus on larger, grammatical aspects. For example, a few of the categories from Underwood & Jongejan are multiple coordination and structural ambiguity and a few categories from Campbell are complex noun phrases and abstractness. These are not enough to classify all the difficult items in this experiment. Since there is not yet a consensus about the indicators of translatability, maybe classifications of other researchers might have been more in line with the types of difficult items in this experiment.

Some recurring reasons the students gave for indicating certain difficult items are terminology, and realia or names, for example for the difficult items 'four counts of murder' and 'International Criminal Court in The Hague' respectively. The students especially seemed to focus on not knowing the exact correct translation of a certain word (group) and did not indicate many grammatical problems. The teachers saw more problems with ing-forms, idiomatic language and typical English structures that cannot be translated literally into Dutch. Examples of these categories are 'fighting flaring up again, 'had been acting' and 'the awareness of other hospital staff'. These examples are only a selection, though these explanations seem to be recurring frequently with students and teachers respectively.

5.3. Process data

This section presents the process data obtained during the translation of text 1 and text 3 by the students, namely duration and number of edits.

5.3.1. Duration

In this section we summarize the segment translation duration information that is available in the CRITT TPR-DB segment table of our experiments. The duration is expressed in milliseconds.

Text	Average Duration per token
T01	4149.5
T03	6663.6

Table 18 - Average duration per token (in milliseconds)

Table 18 shows the average duration per token for the two texts for the entire session. Text 3 took 50% longer to translate than text 1. This already gives a first indication of the general difficulty of the two texts.

Figure 1 shows the normalised translation duration per student in a box-whisker plot. The dots represent outliers, i.e. data points that exceed a distance of 1.5 times the IQR below the first quartile Q1 or 1.5 times the IQR above the third quartile Q3. The high outliers in Figure 1 indicate problematic segments that took a significantly long duration, i.e. where the translation was much slower than the

median for that student. By analysing the specific dataset, we can find the specific segments that correspond to these outliers and they are marked with the call-out shapes in the figure. The outliers are listed in Table 19. For completeness, we also give the median and quartile values in Table 20. A full overview of the normalised translation durations per student per segment are included in appendix H.

In the graph it can be seen that the raw data are indeed normalised: the average is always near 1.00, as indicated by a cross in the box-whisker plot. More specifically, the average values range from 0.88 to 1.11. However, the boxes, whiskers and outliers vary substantially across the group of students. A number of them have a small variation, namely P05 and P08. These students did not spend much more time on difficult segments as compared to their median translation duration. P04 also does not have a big variation, except for the three outliers. Others, such as P01 and P06, spent three or even four times as much time on some segments as compared to their median value.

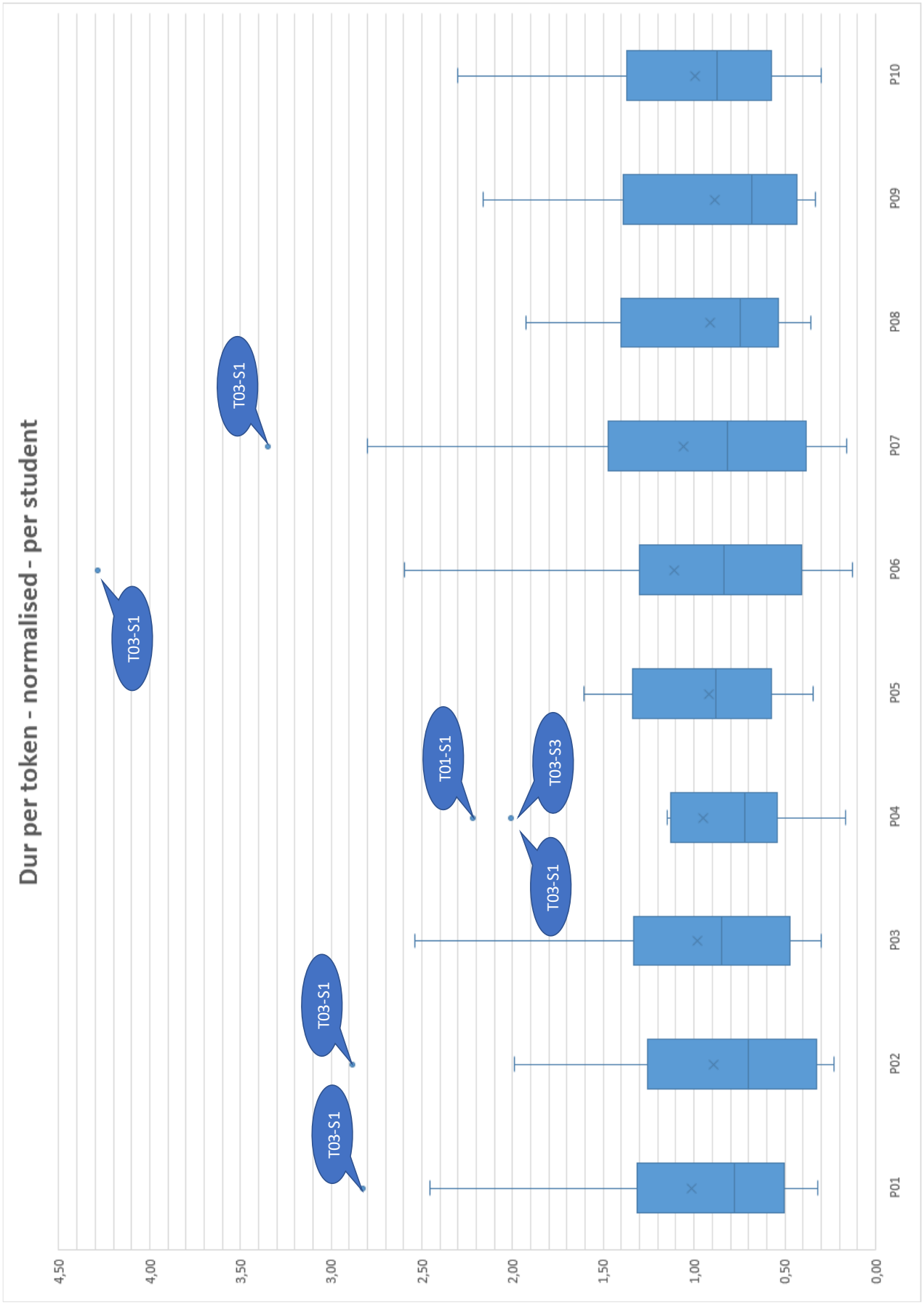

Figure 1 - Normalised duration per student

	Outliers	Normalised Dur per token
P01	T03-S1	2.83
P02	T03-S1	2.88
P04	T01-S1	2.22
	T03-S1	2.01
	T03-S3	2.01
P06	T03-S1	4.29
P07	T03-S1	3.35

Table 19 – Outliers

Part	OutL	Q1	Q2	Q3	OutH
P01	-0.72	0.50	0.78	1.31	2.53
P02	-1.07	0.33	0.70	1.26	2.65
P03	-0.81	0.47	0.85	1.33	2.61
P04	-0.33	0.54	0.72	1.13	2.00
P05	-0.57	0.57	0.88	1.34	2.48
P06	-0.93	0.41	0.84	1.30	2.64
P07	-1.26	0.38	0.81	1.48	3.12
P08	-0.78	0.53	0.74	1.41	2.71
P09	-1.01	0.43	0.68	1.39	2.83
P10	-0.62	0.57	0.87	1.37	2.56

Table 20 - Median and quartile values

In Table 19 we see that five of the ten students had at least one segment that required much more time to translate. One student, P04, has three outliers or three segments for which they needed much more time than for the other segments. In total there are seven outliers for the two texts together, of which three are from the same student. The segments that form outliers are segment 1 from T01, segment 1 from T03 and segment 3 from T03. Five different students took much longer to translate segment 1 from T03, which is the title of the text. Segment 1 from T01 and segment 3 from T03 are only for one student considered outliers, namely for P04.

Figure 2 shows the same data as Figure 1, namely the normalised translation duration, but now the data is grouped per segment. The first 11 boxes belong to T01 and the last 5 to T03. This figure shows visually that T03 took more time to translate as the boxes are generally higher than those of T01. What this figure also shows is that in both texts the first segment took the longest to translate. In both texts this is a title, which seems to indicate that students have the most difficulty with titles.

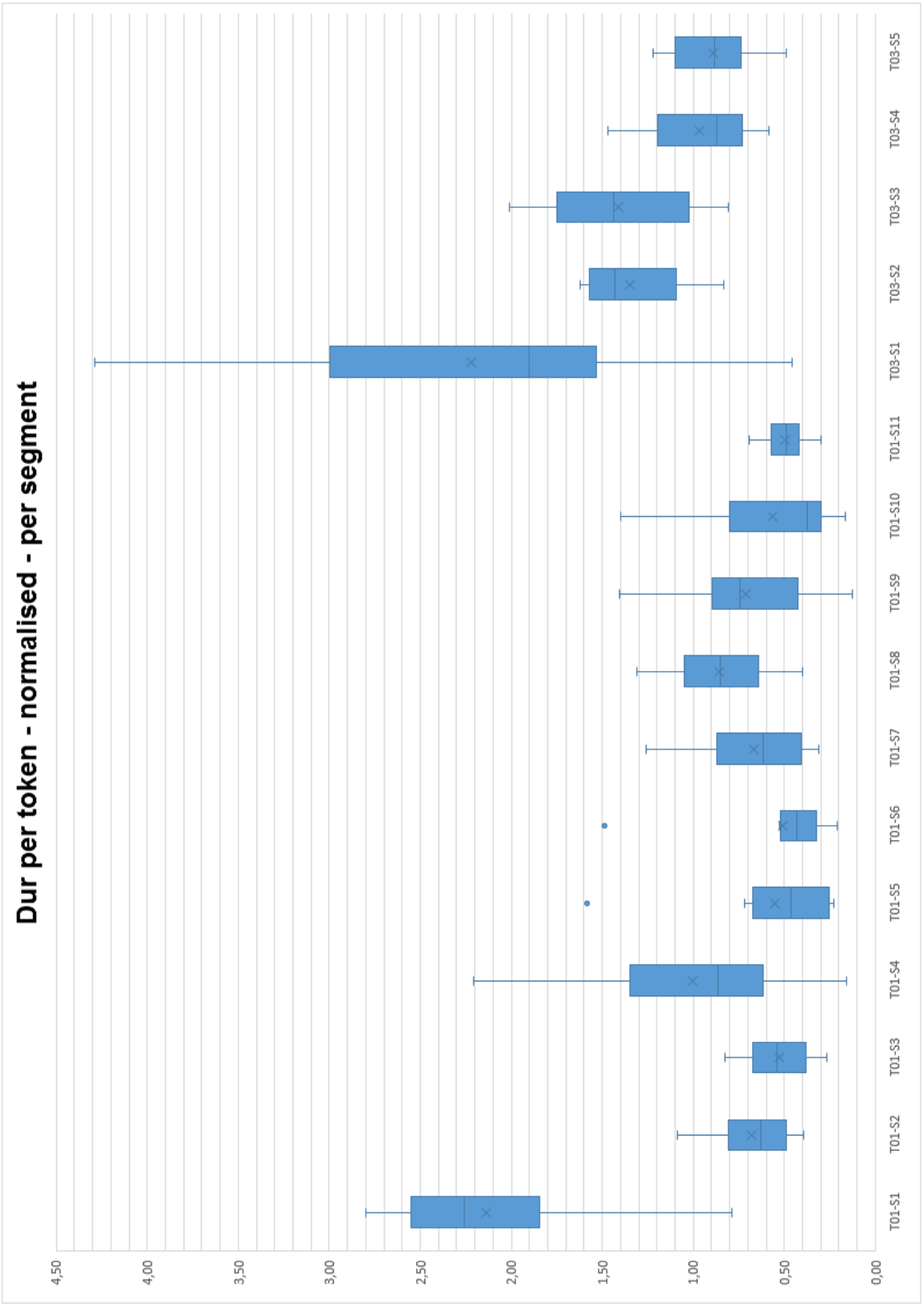

Figure 2 - Normalised duration per segment

		Class						
Text	Segment	1	2	3	4	5	6	Average
T01	1				1	8	1	5
	2		2	5	3			3.1
	3		5	4	1			2.6
	4		1	3	4	2		3.7
	5		5	4		1		2.7
	6		6	3		1		2.6
	7		3	2	5			3.2
	8		1	3	5	1		3.6
	9		3	4	2	1		3.1
	10		7	1	1	1		2.6
	11		5	5				2.5
T03	1		1			4	5	5.2
	2			1	2	7		4.6
	3				4	5	1	4.7
	4			3	5	2		3.9
	5		1	2	7			3.6

Table 21 - Classification normalised duration

The analysis of the normalised duration visualised in the box-whisker plots can also be represented with numbers. As explained in section 4.4.3.1 above, each duration value can be classified based on the quartile to which it belongs, ranging from class 1 to 6. Class 1 and 6 represent low and high outliers respectively and the classes in between, 2 to 5, range from 'lower than Q1, but no outlier' to 'higher than Q3, but no outlier'. For each segment we counted how many students are in each class. The more students in a high class, the higher the normalised duration or the slower the translation of this particular segment and vice versa for the lower class numbers. Table 21 gives the count for all segments of both texts. It gives a general indication of which segments were translated quicker and which slower. The table shows that there are no low outliers for any of the students, but there are in total seven high outliers, namely one time segment 1 from T01, five times segment 1 from T03 and one time segment 3 from T03. Segment 1 of T01 took longer, considering that there are eight students in class 5 and 1 in class 6. Segment 11 of T01 seems to have taken less time, with all students classifying below the median value. In T03 the most difficult segment seems to be segment 1, although one student belongs to one of the lowest classes, namely class 2, which means that this student translated this segment much faster than the median. The easiest segment from T03 is not that easy to identify in this table, because for the other segments the classes are rather varying.

Table 21 also gives the average classification for each segment in the last column. This can be considered a single normalised metric of the duration. Segment 1 of T01 took the longest to translate of all the segments in T01, with an average classification of 5. This means that on average the first

segment of T01, which is the title, took the students longer to translate than their median value. In T03, the segment that took the longest to translate is segment 1, which has an average classification of 5.2. The classification of T03 is slightly higher than segment 1 of T01, which was the segment that required the most time in T01. The last segment of text 3 has an average classification of 3.6 and is the segment that was translate the fastest in T03. In T01, the segment that required the least time is also the last segment, namely segment 11. Additionally, segment 11 is also the segment that was translated the fastest of all the segments in both texts.

5.3.2. Number of edits

We also examined the feature Nedit. This feature shows the number of times a segment was edited. The default value is 1, which means that a participant only once worked on the translation of that segment and did not return to the segment to make changes or add something to the translation. A value higher than 1 indicates that a participant translated a segment and then returned at least once to the segment. It would be expected that the higher the Nedit value, the more difficult the segment is to translate. We have based our analysis on the same dataset, namely the segment table.

T01	Total Nedits	Average Nedits
P01	16	1.5
P02	16	1.5
P03	23	2.1
P04	23	2.1
P05	13	1.2
P06	19	1.7
P07	20	1.8
P08	14	1.3
P09	13	1.2
P10	19	1.7

Table 22 - Number of edits per student - T01

T03	Total Nedits	Average Nedits
P01	5	1
P02	15	3
P03	10	2
P04	8	1.6
P05	9	1.8
P06	21	4.2
P07	20	4
P08	8	1.6
P09	10	2
P10	11	2.2

Table 23 - Number of edits per student - T03

Table 22 and Table 23 show the total and average number of edits per student for text 1 and 3. In text 1, P05 and P09 made the least number of edits, namely 13 edits, which is an average of 1.2 edits per segment. Students 3 and 4 have the most edits, namely 23. This amounts to an average of 2.1 edits per segment. In text 3, P01 made the least number of edits, which is 5 edits for the full text, which equals to one edit per segment. This means that for each segment P01 translated the segment, but did not return to the segment to make a change to the translation. The most number of edits in T03 is 21 and is from student 6. The average for student 6 is 4.2 edits per segment.

It is also important to examine which segments took several revision cycles and which segments took more cycles than others. A summary of this information is shown in Table 24. It shows the count of the number of students that required a certain number of revisions.

		Nedit								
Text	Segment	1	2	3	4	5	6	7	Average Nedits	Total Nedits
T01	1	3	5	1			1		2.2	22
	2	3	5	1	1				2	20
	3	7	2	1					1.4	14
	4	5	4			1			1.8	18
	5	6	2	2					1.6	16
	6	8	2						1.2	12
	7	5	4	1					1.6	16
	8	6	3	1					1.5	15
	9	6	4						1.4	14
	10	5	4	1					1.6	16
	11	7	3						1.3	13
T03	1	5	2	2	1				1.9	19
	2	2	2	4	1			1	2.9	29
	3	2	4	2	1			1	2.7	27
	4	2	6		1		1		2.4	24
	5	3	6	1					1.8	18

Table 24 - Count of Nedit

For some segments it took some students 4, 5 or even up to 7 edits to translate the segment. This is the case, for example, for segment 1 of T01 and segment 2 of T03. For other segments, all except a few students required only one edit, for example for segment 6 of T01, where 8 of the 10 students only needed one edit. This suggests which segments are difficult to translate and which are easier.

Table 24 above also shows per segment for both texts the total number of edits (Nedits) and the average number of edits per student. The lowest Nedits in T01 is 12 edits for segment 6 or an average of 1.2 edits per student. For T03 segment 5 was edited the least number of times, being edited 18 times. The segments that were edited the most often are segment 1 for text 1, with 22 edits, and segment 2 in text 3, with 29 edits or 2.9 edits per student.

6. Discussion

In this section we evaluate the results of our experiment in light of the research questions given in section 3 and we formulate some conclusions. First, the relationship between readability and translatability is covered. Then we look at the subjective rating of difficult textual elements to see if there are commonalities between the different participants. Finally, we evaluate the correlation between the process data, i.e. duration and number of edits, and the subjective rating of the participants.

6.1. Readability and translatability

The first research question aims to see whether the readability of a text is a reliable feature to predict the translatability of this text. In order to find an answer we compare the readability of the two texts with the subjective rating by both teachers and students, and also with the translation duration needed to translate the texts.

We compared the subjective ratings of the texts with the readability measure conducted in advance. This showed that both teachers and students rated the texts similarly to the readability results. All students and teachers, except one teacher, deemed text 3 to be more difficult than text 1, which is also reflected in the readability scores that suggested that text 3 was the most difficult text of all six texts of the original dataset. The difference between the two texts is not that big, but does show that text 3 is somewhat more difficult to translate not only according to teachers, but also according to students. There was, however, a larger difference in the classification of the two texts by the students as compared to the ratings from the teachers. This means that generally the students found text 1 easier and text 3 more difficult than the teachers. This might be explained by the fact that the students first actually translated both texts before rating them, while the teachers did not translate the texts, but had to estimate the difficulty for students. Another explanation could be that students may consider that something is easy to translate, while it is actually difficult. Students may not be aware that these items are difficult to translate and in fact make more errors in the translation. It should also be noted that the question for the students was not exactly the same as for the teachers. The teachers had to indicate difficult parts of the texts, while students were asked to indicate what they would want to look up, which may not be exactly equivalent to translation difficulty. Therefore they might have marked items that are easier to look up in external resources. The teachers therefore focused on different aspects of difficulty than the students. Maybe if the question was phrased differently for the students, they might have indicated slightly different items.

The general conclusion is similar to one of the conclusions of the research of Liu, Zheng & Zhou (2019). They could conclude that the subjective rating of translatability was in line with the readability of the

texts involved and in line with the process data. In our experiment, the teachers and the students also rated the two texts in accordance with the readability measures.

We find that not only the subjective assessment confirms the relationship between readability and translatability, but also the process data supports the classification of translation difficulty. The average translation duration of the participants in this experiment shows that text 3 took longer to translate than text 1, which could also be deduced from the average translation duration for translation into other languages that was used to estimate the readability. A text that is rated as more difficult to translate also takes a longer time to translate and requires more revisions. The box-whisker plot shows that text 3 is generally more difficult, because six of the seven outliers are from segments in text 3. Additionally, if you look at the average translation duration per text, text 1 is overall translated faster than the median value and text 3 takes longer to translate than the median value. These results are in line with the readability measures that rated text 3 as more difficult to read. The Nedits also show that text 3 is more difficult, because on average more edits were made in text 3. Therefore, readability can be considered a reliable indicator of translation difficulty and can be used to give a general indication of the translatability of a text before actually translating the text.

6.2. Agreement in difficult items

The second research question addresses the subjective ratings: have teachers and students marked the same or similar textual elements (sentences, word groups or individual words) as being difficult? Is there a substantial difference between teachers and students? And can we link the items to indicators found in the research literature? On this subject, the results are not very conclusive. Nevertheless, we can summarize some general observations.

Figure 3 shows the correlation between how many difficult items were indicated by teachers and students in text 1. Figure 4 shows the same information for text 3. In both graphs the x-axis shows the percentage of difficult items that teachers indicated in each segment and the y-axis shows the percentage of difficult items indicated by the students in each segment. The figures were developed based on Table 25, of which each row represents one dot in the figures. Figure 3 and Figure 4 show that there is only a weak correlation between the segments that were considered most difficult by the teachers and by the students. The correlation for text 1 is 0.44 and that for text 3 is 0.50. This shows that the teachers and students are not completely in agreement about which items are difficult to translate. For some items there is more agreement than others, but overall there is not a lot of agreement between the teachers and the students. There is, however, a little bit more agreement for text 3, although the difference with text 1 is not that big.

T01	Teachers	Students
1	16.22%	7.89%
2	9.46%	18.42%
3	6.76%	10.53%
4	16.22%	23.68%
5	4.05%	10.53%
6	4.05%	2.63%
7	10.81%	10.53%
8	12.16%	5.26%
9	12.16%	2.63%
10	4.05%	5.26%
11	4.05%	2.63%
Correlation:		**0.44**
T03		
1	14.29%	10.61%
2	16.67%	19.70%
3	32.14%	22.73%
4	17.86%	22.73%
5	19.05%	24.24%
Correlation:		**0.50**

Table 25 - Correlation difficult items - segment level

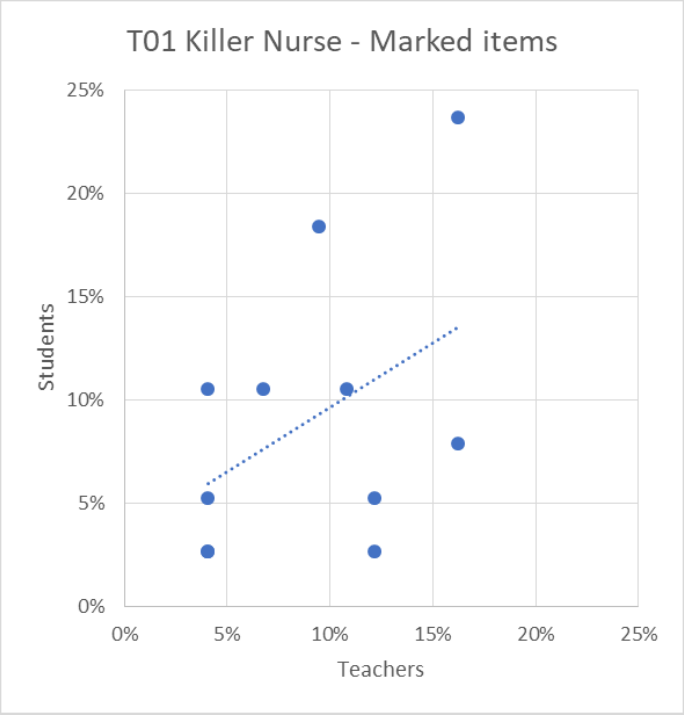

Figure 3 - Correlation difficult items on segment level - T01

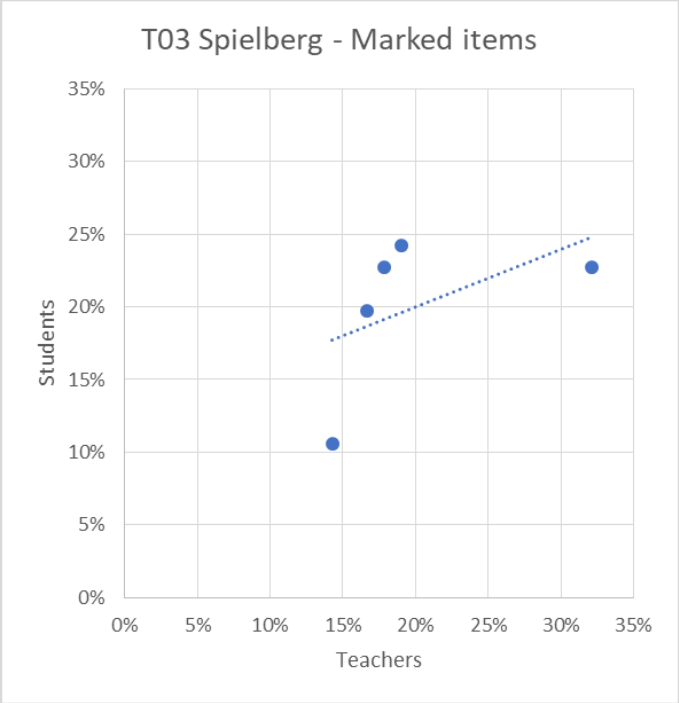

Figure 4 - Correlation difficult items on segment level - T03

On the item level, there is also a difference in the number of marked items between both groups of participants. In both texts the teachers indicated more difficult items than the students. As already mentioned in the previous section, a possible reason for this might be because the teachers and students received a slightly different question. However, there is also a difference amongst the teachers about the items that were considered difficult. This is shown by the most frequently indicated items. In both texts there are only four of the 28 different difficult items in T01 and T03 that were indicated by half of the teachers. There is also no difficult item that was indicated by more than half of the teachers in T01. For the students, there are only two difficult items of the 16 in T01 that were marked by at least half of the students. In text 3 this number is 5, of a total of 27 different difficult items. There is not one difficult item that is indicated by all the teachers or all the students or both.

This conclusion is similar to the findings of Eyckmans & Anckaert (2017). Their research showed that for two different assessment methods, namely the CDI and PIE method, there was little agreement between the assessors when indicating difficult items. This is also the case in our experiment: not only is there little agreement among the teachers, but the correlation between the teachers and students is also low. Eyckmans & Anckaert conclude then that in order to objectively asses a translation, multiple assessors are needed. Our experiment, which involved nine teachers who could be considered as assessors, seems to confirm this.

There also seems to be a difference between teachers and students when we look at what types of items have been indicated as difficult. Students focused more on the specific translation of words in

their context, and on names. Teachers, however, focused more on grammatical difficulties, such as ing-forms and tenses, and idiomatic language. One explanation for this is that the two groups of participants received a slightly different prompt. The teachers were asked to indicate things that they thought students would have difficulties with, whereas students were asked to indicate what they would have looked up if they could have used resources. The students may have focused more on the translation of the words and how to correctly translate and spell names, and not on grammatical problems, because it is easier to look up translations in external resources than solutions to grammatical problems in a specific context.

6.3. Translation process

The final research question concerns the relationship between translatability and the translation process, for which process data was collected. Specifically, we focused in this study on data about the duration in the session and segment table and number of edits available in the segment table. The final research question also aims to see if the process data can be linked to the difficulties indicated by the teachers and students.

The data about the translation duration shows that T03 is more difficult to translate than T01. After normalising the durations, almost all the outliers belong to T03. When looking at the average classification of T01 and T03, T03 is also clearly indicated as more difficult, considering that the segments of T01 took on average less time than the median to be translated, while the segments from T03 took on average longer than the median. The average normalised duration for T03 is 72% higher than for T01: 0.79 for T01 versus 1.37 for T03.

The other process feature of focus is the number of edits. This feature also showed that T03 is more difficult than T01. The segments in T03 were edited, on average, more times than the segments in T01. Concretely, on average each participant made 1.6 edits per segment in T01, while 2.2 edits were made in T03.

These findings are in line with the findings of Vanroy, De Clercq & Macken (2019) and Bangalore et al. (2015). Both studies found that revision and duration features are a reliable indicator of the translatability of a text. Our experiment shows the same results, since both the duration and the number of edits show that text 3 is the most difficult text to translate out of the two texts. The results of this experiment contradict the result of Hvelplund (2011) who found that a difficult text does not require more cognitive effort, but he used the feature pupil size, and this may not be a reliable feature to measure cognitive load, like Liu, Zheng & Zhou (2019) and Hvelplund proposed.

The process data analysed in this study clearly shows that T03 is the most difficult text and T01 is the easiest. This is in line with the overall subjective ratings from the teachers and the students and with

the readability measures. Therefore, the translation duration and the number of edits can be considered a good indicator of the translatability of a text, as well as the subjective rating of the translatability. This supports the conclusion of Liu, Zheng & Zhou (2019). They found that subjective ratings can give a general indication of the translatability of a text, because the subjective rating of the translatability of students was in line with the process data. Process data can then show in more detail which specific elements in the text are more difficult.

In section 5.2.1 we examined which specific segments teachers and students found more difficult to translate. Even though the overall rating of the texts agrees with the process data, there is some difference in which segments are considered more difficult. The teachers indicated segment 1 and 4 in T01 and segment 3 in T03 as most difficult. For the students in T01 the most difficult segment is 4 and in T03 the most difficult segments are 3, 4 and 5. However, the translation duration data show that segment 1 of T03 is the most difficult segment over the two texts together. This segment did not appear to be the most difficult for the teachers nor for the students. Other outliers were segment 1 from T01 and segment 3 from T03, although according to the process data these segments were only difficult for one student. Both these segments were considered the most difficult in the subjective evaluation, by the teachers and the students respectively. When we compare the subjective evaluation of the segments with the number of edits, we reach a similar conclusion, namely no direct one-to-one relationship.

In Table 26 we summarise the information from our work in a single table. The table contains the number of difficult items identified by the students, the average duration classification and the number of edits. The percentage of difficult items is different here than in Table 10 and Table 12, because in order to be able to measure the correlation over both texts, the percentages were calculated over the two texts. For the correlation we have also not included the percentages of the teachers. Since the process data, namely the duration and number of edits, pertains to the students, it seemed more logical to compare it only with the difficult items the students indicated. The mathematical correlation between the number of difficult items and duration classification is 0.52. This is not very strong, but at least suggest that there is “some” correlation.

Text	Segment	Number of difficult Items	Percentage	Average duration classification	Nedit
T01	1	3	2.88%	5	2.2
	2	7	6.73%	3.1	2
	3	4	3.85%	2.6	1.4
	4	9	8.65%	3.7	1.8
	5	4	3.85%	2.7	1.6
	6	1	0.96%	2.6	1.2
	7	4	3.85%	3.2	1.6
	8	2	1.92%	3.6	1.5
	9	1	0.96%	3.1	1.4
	10	2	1.92%	2.6	1.6
	11	1	0.96%	2.5	1.3
T03	1	7	6.73%	5.2	1.9
	2	13	12.50%	4.6	2.9
	3	15	14.42%	4.7	2.7
	4	15	14.42%	3.9	2.4
	5	16	15.38%	3.6	1.8
			Correlation:	**0.52**	**0.77**

Table 26 - Correlation difficult items on segment level & process data

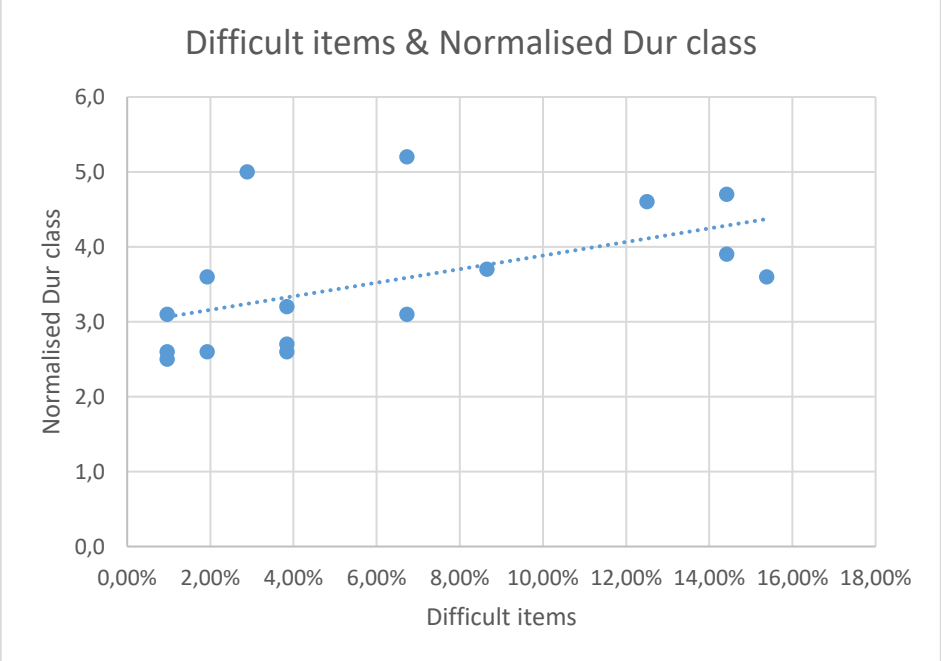

Figure 5 - Correlation difficult items & normalised duration

Figure 5 shows the same data in a graphical format. This visual graph gives a stronger support for the statement that there is indeed a relationship. The two extreme points at the top are responsible for

the low correlation value, and these two correspond to the first segments of both texts, which are the titles. This means that the participants spent a relatively long time on translating the title while it was considered average difficulty. Since translating a title is indeed always an important part of the translation activity, it can be expected that a student contemplates a bit longer than necessary, even though the translation itself is maybe not necessarily much more difficult. Another possible explanation is that titles often contain more creative language, such as a metaphor or a pun, which is often difficult to correctly translate into the target language. As a thought experiment, we calculated the correlation when the title is excluded, and then it is 0.80.

Table 26 also shows that the correlation between the number of difficult items and the number of edits is 0.77, a better correlation than for the duration. The corresponding graphical representation is given in Figure 6, which visually confirms the same conclusion, namely that there is a relationship between difficult items and number of edits. It might also be interesting to calculate the correlation between the duration and the number of edits. The correlation between these two process features is 0.75, which is similar to the correlation between the difficult items and the number of edits. The correlation between the duration and Nedits is also shown in a graph, Figure 7. This means that if a student was slower in translating a text, they also made more edits to the translation.

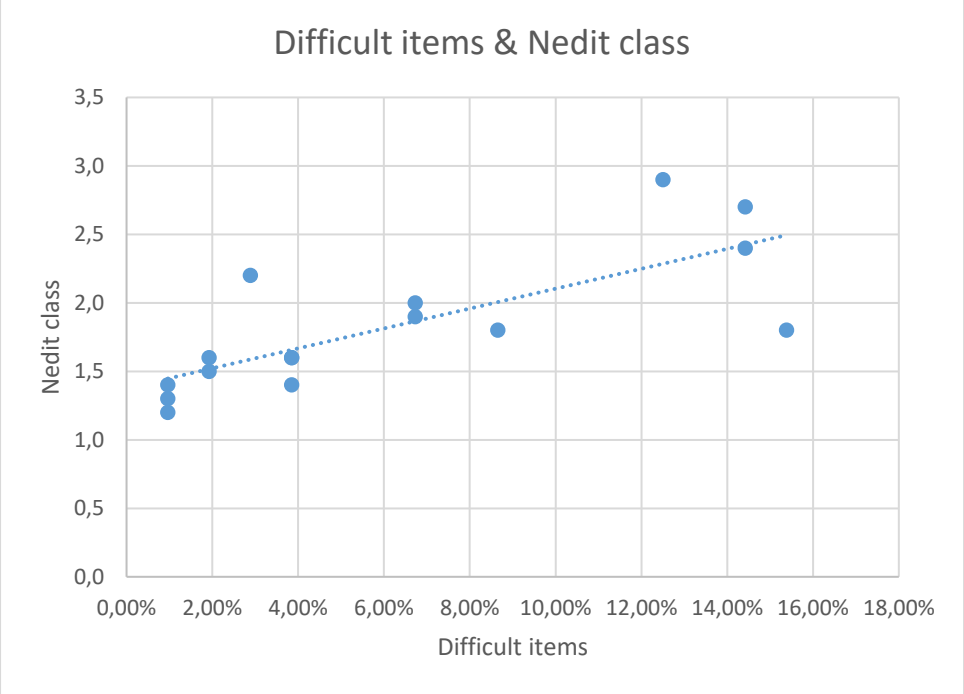

Figure 6 - Correlation difficult items & number of edits

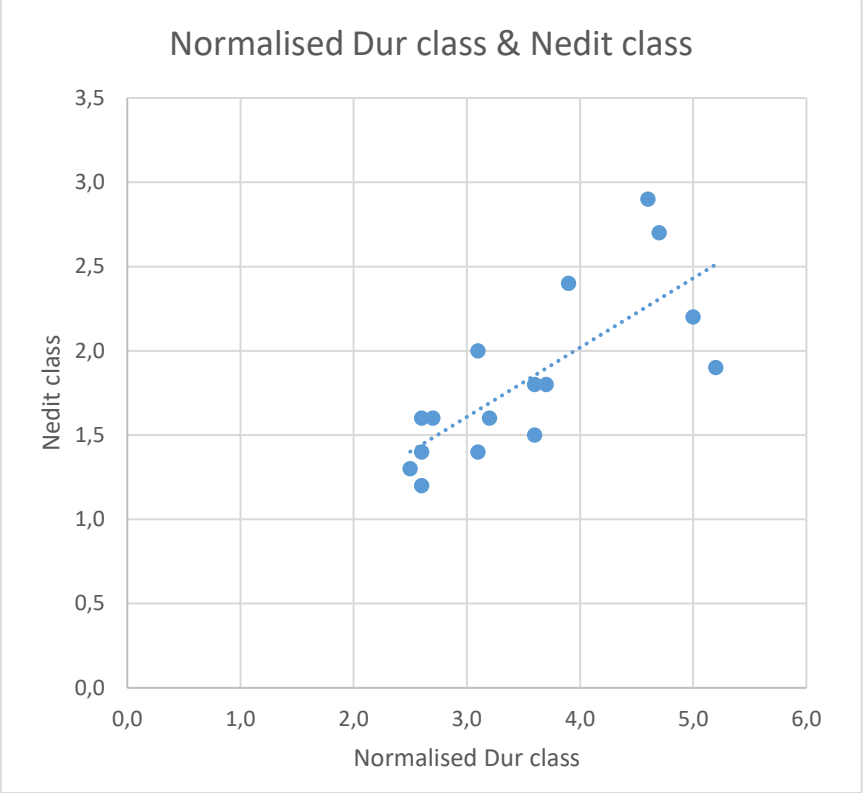

Figure 7 - Correlation normalised duration & number of edits

As a final topic we compare the language proficiency of the students, as measured with the LexTALE test, with the process data. Table 27 shows the results of the LexTALE test per participant, if the participants had any outliers as shown in Figure 1 and the average translation duration per token for the two texts together, calculated using the session duration. The average in this table is not the normalised average, because we want to see the differences between the students. When looking at the outliers, there does not seem to be a link between having an outlier and the language proficiency of that student, since one of the participants with the lowest proficiency, namely P08 with 87.5%, did not have an outlier, while P04, who scored 95%, has three outliers. When comparing the proficiency level of the students with the average duration they needed to translate both texts, there also does not seem to be a correlation. We should remark that all the students study English translation and therefore all have a high proficiency of English. According to LexTALE, the average score for Dutch advanced learners of English is 70.7%. All the students in this study have higher scores than the average. Therefore we could not compare participants with a big difference in proficiency. The data, however, can give a general indication of whether there might be a correlation between proficiency and cognitive effort. P10 has the highest proficiency score, namely 100%, but not the fastest time. The student who translated the fastest is P08, but has one of the lowest LexTALE results, 87.50%, although this is still a very high score. To conclude, we can see that the level of proficiency does not seem to

have an influence on the cognitive effort required to translate a text. The correlation between the duration and LexTALE is only 0.29, showing that there is no real relationship between the language proficiency of students and the process data.

Participant	LexTALE results	Outliers	DurT
P01	87.50%	1	3512.61
P02	97.50%	1	5889.56
P03	92.50%	0	5436.43
P04	95%	3	9181.17
P05	98.75%	0	3087.36
P06	92.50%	1	7123.11
P07	97.50%	1	7628.11
P08	87.50%	0	2357.18
P09	97.50%	0	4501.38
P10	100%	0	4773.8

Table 27 - LexTALE results, outliers & average duration per token (in milliseconds)

7. Future work

In our work we have investigated the task of assessing translation difficulty based on readability, textual indicators and process data. The investigation led to some conclusions, but some results are inconclusive and not all questions have been answered completely. One general suggestion for future research is to conduct the same experiment, but with more data. Since our experiment is based on a relatively small dataset, it would be useful to conduct the same experiment with more data to see if it leads to similar results. Below we propose a few more possible topics of future research.

7.1. Indicators

As explained earlier, our experiment could not confirm major indicators of translation difficulty. We expected to find that teachers and students would, to some extent, indicate common indicators in the texts and that some of these indicators would confirm the results of the research literature. However, the results were not as simple. In our experiment the participants were allowed to describe the reason for marking an item as being difficult in their own words. It was therefore sometimes difficult to interpret the results: for example, did a person identify with his textual comment a certain indicator as described in the literature, did person A and person B mean the same thing with their comments?

One might suggest to adjust the experiment by asking the participants to select one indicator of a set of possible indicators. For example, a list of the 10 most common indicators found in the current literature could be compiled. This would certainly make the processing easier and the interpretation of the results more objective. However, such an approach also introduces the risk of biasing the results towards these indicators. So at least the set of indicators presented to the participants should include also an option 'other'. If the latter option is used frequently then we can conclude that the selected indicators for the experiment are not a good choice and the selection should be revised. If the 'other' option is not used often then we can conclude that the indicators chosen by the participant are really what they meant.

7.2. Process Data

Our experiment was done using the TransLog program, and this resulted in a lot of data at different levels (see https://sites.google.com/site/centretranslationinnovation/tpr-db/features for a description of all the data). We restricted our work to only two features: duration, from the segment and session tables, and number of revisions, from the segment table. Obviously, other data is also likely relevant for the analysis of translation difficulty, and not only from the session and segment tables. For instance, pause duration and number of deletions also seem relevant for a translation activity.

In fact, one important set of data that has not been part of our work, although initially we had planned to use it, is eye tracking data. When we started our work it was planned to also collect eye tracking data. Unfortunately, the Covid-19 measures did not allow students to come on-site for the eye tracking experiment. Eye tracking information, however, is certainly very meaningful, because it allows to correlate what part of the source text is being looked at for how long by the participant when working on what part of the translation. This is certainly an area worth investigating in future work.

7.3. Texts

Finally, one can challenge the selection of the two texts. They are short texts that have been used in a lot of other research. However, it is important to evaluate if the results are also applicable to longer texts that are common in the translation profession, and other types of texts, such as legal documents, business texts, and webpages.

8. Conclusion

The aim of this master thesis was to estimate the translatability of two English texts based on readability scores, subjective evaluation and translation process data. Firstly, we used readability scores to classify the texts, since it is assumed that the readability of a text also gives an indication about the translatability of that text. After examining the readability, we looked at the subjective rating that teachers of English translation and translation students gave to the two texts. Next, we compared which items of the texts were considered difficult by the teachers and students. Lastly, we examined the translation process data acquired by logging the keyboard activity of the students when translating the two texts. The specific process features we focused on were duration and number of edits.

We could conclude that readability is indeed a good indicator of the translatability of a text. There is agreement between the readability scores of the two texts and the subjective ratings on the one hand and between the readability scores and the process data on the other hand.

When looking closer at the difficult items indicated by the teachers and students, we found that there was not that much correlation between the teachers and the students. Both groups indicated different items and focused on different types of difficulty in a text. Even though both groups of participants agreed that text 1 is easier and text 3 is more difficult, the reasons for this are different.

We also examined the process data and found that both process features in this study, i.e. translation duration and number of edits, are reliable indicators of translation difficulty. The results showed that students who translated slower also made more edits to their translation. These features also agree with the readability scores and the ratings from the participants. However, the segments of the texts marked by the participants as most difficult did not correlate well with the process data. The most difficult segments according to the teachers and students did not appear to be the segment that was most often an outlier. The number of edits had a stronger correlation with the difficult items than with the duration. The correlation with duration is influenced by the titles of the texts: if we leave out these segments, then this results in a better correlation.

Finally, we could not find a correlation between the language proficiency of the students and the process data. However, the level of language proficiency of all the participants was quite alike; no one has a very low proficiency of English, because they are students of English translation. Therefore the correlation with language proficiency is not clear.

9. Bibliography

Bangalore, S., Behrens, B., Carl, M., Ghankot, M., Heilmann, A., Nitzke, J., Schaeffer, M. & Sturm, A. (2015). The role of syntactic variation in translation and post-editing. *Translation Spaces, 4* (1), 119-143.

Campbell, S. (1999). A Cognitive Approach to Source Text Difficulty in Translation. *Target, 11* (1), 33-63.

Campbell, S. (2000). Choice Network Analysis in Translation Research. In M. Olohan (Ed.), *Intercultural Faultlines* (pp. 29-42). London: Routledge.

Collins-Thompson, K. (2014). Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL - International Journal of Applied Linguistics, 165* (2), 97-135. doi: 10.1075/itl.165.2.01col

CRITT, Centre for Research and Innovation in Translation and Translation Technology. (s.d.). Retrieved from https://sites.google.com/site/centretranslationinnovation/tpr-db.

Dale, E. & Chall, J.S. (1949). The Concept of Readability. *Elementary English*, *26* (1), 19-26. Retrieved from https://www.jstor.org/stable/41383594

De Clercq, O. & Hoste, V. (2016). All Mixed Up? Finding the Optimal Feature Set for General Readability Prediction and its Application to English and Dutch. *Computational Linguistics*, *42* (3), 457-490. doi: 10.1162/COLI_a_00255

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M. & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, *20* (3), 293-325. doi: 10.1017/S1351324912000344

DuBay, W.H. (2004). *The Principles of Readability*. Costa Mesa, CA: Impact Information.

Eyckmans, J., & Anckaert, Ph. (2017). Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*, 40-56.

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology, 32* (3), 221-233. doi: 10.1037/h0057532

Hvelplund, K.T. (2011). *Allocation of Cognitive Resources in Translation: an eye-tracking and key-logging study.* Copenhagen: Copenhagen Business School.

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye Movement Behaviour Across Four Different Types of Reading Task. *Copenhagen Studies in Language, 36*, 103-124.

Jensen, K.T.H. (2009). Indicators of text complexity. *Copenhagen Studies in Language,* (37), 61-80.

Kraf, R. & Pander Maat, H. (2009). Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing, 31* (2), 97-123.

Liu, Y., Zheng, B. & Zhou, H. (2019). Measuring the difficulty of text translation. The combination of text-focused and translator-oriented approaches. *Target. International Journal of Translation Studies, 31* (1), 125 - 149. Retrieved from https://www.jbe-platform.com/content/journals/10.1075/target.18036.zhe

McLaughlin, G. H. (1969). SMOG Grading – a New Readability Formula. *Journal of Reading, 12* (8), 639-646. Retrieved from https://www.jstor.org/stable/40011226

McLaughlin, G. H. (1974). Temptations of the Flesch. *Instructional Science, 2* (4), 367-384. Retrieved from https://www.jstor.org/stable/23368030

Sun, S. (2015). Measuring translation difficulty: theoretical and methodological considerations. *Across Languages and Cultures, 16* (1), 29-54. doi: 10.1556/084.2015.16.1.2

Underwood, N.L. & Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT. *Proceedings of MT Summit VII: Machine Translation in the Information Age*, 363-368. Retrieved from http://www.mt-archive.info/MTS-2001-Underwood.pdf

Vandepitte, S. (2016). *Translating Untranslatability – English - Dutch / Dutch - English*. Gent: Academia Press.

Vanroy, B., De Clercq, O., & Macken, L. (2019). Correlating process and product data to get an insight into translation difficulty. *Perspectives, 27* (6), 924-941.

Appendix A: T01

1 Killer Nurse Receives Four Life Sentences

2 Hospital nurse Colin Norris was imprisoned for life today for the killing of four of his patients.

3 32 year old Norris from Glasgow killed the four women in 2002 by giving them large amounts of sleeping medicine.

4 Yesterday, he was found guilty of four counts of murder following a long trial.

5 He was given four life sentences, one for each of the killings.

6 He will have to serve at least 30 years.

7 Police officer Chris Gregg said that Norris had been acting strangely around the hospital.

8 Only the awareness of other hospital staff put a stop to him and to the killings.

9 The police have learned that the motive for the killings was that Norris disliked working 10 with old people.

10 All of his victims were old weak women with heart problems.

11 All of them could be considered a burden to hospital staff.

Appendix B: T03

1 Spielberg Shows Beijing Red Card over Darfur

2 In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China's backing for Sudan's policy in Darfur.

3 His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government.

4 China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague.

5 Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community, and particularly China, should do more to end the suffering.

Appendix C: T04

Climate Change

Although developing countries are understandably reluctant to compromise their chances of achieving better standards of living for the poor, action on climate change need not threaten economic development. Incentives must be offered to encourage developing countries to go the extra green mile and implement clean technologies, and could also help minimise emissions from deforestation. Some of the most vulnerable countries of the world have contributed the least to climate change, but are bearing the brunt of it. Developing countries, in particular, need to adapt to the effects of climate change. Adaptation and mitigation efforts must therefore go hand in hand.

Appendix D: Questionnaire teachers

Vragenlijst docenten

Masterproef translation difficulty

*Vereist

1. Voornaam *

2. Achternaam *

3. Aan welke instelling(en) geeft u les? *

4. Hoeveel jaar ervaring hebt u als docent? *

Markeer slechts één ovaal.

- 1-5 jaar
- 5-10 jaar
- 10-15 jaar
- 15-20 jaar
- >20 jaar

5. Welke vakken geeft u? *

6. Hoe schat u de moeilijkheid van tekst A in? (Killer Nurse Receives Four Life Sentences) *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	8	9	10	
Gemakkelijk	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Moeilijk

7. In welk jaar zou u tekst A laten vertalen? *

Markeer slechts één ovaal.

- ◯ Eerste bachelor
- ◯ Tweede bachelor
- ◯ Derde bachelor
- ◯ Master

8. Hoe schat u de moeilijkheid van tekst B in? (Spielberg Shows Beijing Red Card over Darfur) *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	8	9	10	
Gemakkelijk	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Moeilijk

9. In welk jaar zou u tekst B laten vertalen? *

Markeer slechts één ovaal.

- Eerste bachelor
- Tweede bachelor
- Derde bachelor
- Master

10. Waarom vindt u tekst A of B moeilijker? Of waarom schat u ze even moeilijk in? *

11. Met welk soort elementen uit de teksten zouden studenten het meeste moeite hebben? Waarom? *

Appendix E: Questionnaire students

Vragenlijst

Masterproef translation difficulty

*Vereist

1. Voornaam *

2. Acternaam *

3. Geslacht *

Markeer slechts één ovaal.

- Man
- Vrouw
- Andere

4. Leeftijd *

5. In welk jaar zit je momenteel? *

Markeer slechts één ovaal.

- Derde bachelor
- Master Vertalen
- Postgraduaat CALM

6. Talencombinatie *

7. Heb je al een masterdiploma behaald? *

Markeer slechts één ovaal.

- Ja
- Nee

8. Welke masterdiploma's heb je al behaald?

9. Aan welke instelling heb je deze masterdiploma's behaald?

10. Met welk soort items uit de teksten had je het meeste moeite? Waarom? *

11. Hoe moeilijk vond je het om tekst A te vertalen? (Killer Nurse Receives Four Life Sentences) *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	8	9	10	
Gemakkelijk	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Moeilijk

12. Hoe moeilijk vond je het om tekst B te vertalen? (Spielberg Shows Beijing Red Card over Darfur) *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	8	9	10	
Gemakkelijk	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Moeilijk

13. Waarom vond je tekst A of B moeilijker te vertalen? Of waarom vond je ze even moeilijk? *

Appendix F: Difficult items – T01

Segment	Difficult item	D01	D02	D03	D04	D05	D06	D07	D08	D09	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	D	P	Difference
1	Killer Nurse		x			x	x	x	x	x	x								x	x	6	3	3
1	receives			x						x											2	0	2
1	four life sentences						x	x	x	x											4	0	4
2	Hospital nurse	x			x								x	x	x	x	x				2	5	-3
2	Colin Norris																		x		0	1	-1
2	was imprisoned			x				x													2	0	2
2	today			x																	1	0	1
2	for the killing of			x							x										1	1	0
2	four of his patients			x																	1	0	1
3	killed	x																			1	0	1
3	the four women			x																	1	0	1
3	in 2002			x																	1	0	1
3	large amounts of										x								x		0	2	-2
3	sleeping medicine			x			x						x			x					2	2	0
4	yesterday			x																	1	0	1
4	found guilty														x	x					0	2	-2
4	four counts of murder	x	x	x	x	x	x		x	x	x	x	x		x	x				x	8	6	2
4	following a long trial			x			x	x					x							x	3	2	1
5	was given			x																	1	0	1
5	four life sentences			x						x	x		x								2	2	0
5	each of the killings													x		x					0	2	-2
6	serve				x				x	x							x				3	1	2
7	Police officer						x		x			x					x			x	2	3	-1
7	Chris Gregg																		x		0	1	-1
7	had been acting	x		x				x	x												4	0	4
7	around the hospital			x			x														2	0	2
8	the awareness of other hospital staff		x	x			x		x	x					x			x			5	2	3
8	put a stop to him and to the killings	x		x					x	x											4	0	4
9	have learned	x		x		x			x	x											5	0	5
9	working with old people			x					x												2	0	2
10	all of his victims			x																	1	0	1
10	heart problems						x											x			1	1	0
11	could be considered a burden				x																1	0	1

Appendix G: Difficult items – T03

Segment	Difficult item	D01	D02	D03	D04	D05	D06	D07	D08	D09	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	D	P	Difference
1	Shows					x	x	x					x								3	1	2
1	Beijing											x									0	1	-1
1	Red Card			x		x							x								2	1	1
1	over	x				x	x														2	0	2
1	Darfur			x					x		x	x	x			x	x		x		2	6	-4
2	In a gesture sure to	x	x						x	x			x								4	1	3
2	rattle				x	x		x	x	x		x				x	x	x			5	4	1
2	Government			x		x	x			x											4	0	4
2	pulled out of														x						0	1	-1
2	Beijing Olympics												x			x				x	0	3	-3
2	China's backing for Sudan's policy		x	x			x		x		x	x									4	2	2
3	withdrawal	x		x											x	x					2	2	0
3	comes								x												1	0	1
3	in the wake of					x	x			x											3	0	3
3	fighting flaring up again	x	x	x			x		x		x		x	x		x	x				5	5	0
3	set to					x			x	x								x			3	0	3
3	embarrass																				0	1	-1
3	China																		x		0	1	-1
3	sought to					x			x	x											3	0	3
3	halt the							x		x				x							2	1	1
3	negative fallout					x		x	x	x				x		x					4	2	2
3	having close ties to		x			x	x		x	x		x	x								5	2	3
3	Sudanese government			x							x								x	x	1	3	-2
4	extensive investments			x			x		x												3	0	3
4	Sudanese oil industry			x																x	1	1	0
4	maintains												x								0	1	-1
4	which includes one minister				x		x		x												3	0	3
4	crimes against humanity											x			x	x			x	x	0	5	-5
4	International Criminal Court in The Hague			x					x	x	x	x	x			x	x		x	x	3	8	-5
5	Although emphasizing	x	x		x	x			x		x	x	x			x	x		x	x	5	0	5
5	Khartoum			x						x	x	x	x					x			2	7	-5
5	bears the bulk of the responsibility			x					x	x			x								3	2	1
5	ongoing atrocities		x					x	x	x							x	x			4	2	2
5	maintains								x	x			x								2	1	1
5	international community										x					x					0	2	-2
5	the suffering			x											x			x			1	2	-1

Appendix H: Normalised duration per token

Text	Segment	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
T01	**1**	2.45	0.79	2.54	2.22	1.61	2.59	2.80	1.93	2.16	2.30
	2	0.50	0.62	1.09	0.61	0.39	0.48	1.05	0.64	0.69	0.73
	3	0.59	0.27	0.55	0.69	0.39	0.39	0.37	0.53	0.83	0.67
	4	0.88	1.99	2.21	0.75	0.86	1.13	0.16	0.66	0.50	0.91
	5	0.32	0.23	1.58	0.72	0.66	0.26	0.24	0.35	0.58	0.58
	6	0.52	0.26	0.53	0.52	0.34	0.21	1.49	0.40	0.46	0.36
	7	0.47	0.51	0.31	0.72	0.94	0.84	0.43	0.85	0.33	1.26
	8	1.05	0.40	0.84	1.05	1.01	1.31	0.87	0.83	0.67	0.57
	9	0.44	0.86	0.74	0.98	0.87	0.13	0.76	0.56	0.40	1.41
	10	1.40	0.30	0.30	0.17	0.65	1.26	0.34	0.39	0.36	0.50
	11	0.69	0.54	0.43	0.44	0.55	0.63	0.42	0.54	0.42	0.30
T03	**1**	2.83	2.88	0.45	2.01	1.39	4.29	3.35	1.58	1.60	1.80
	2	1.56	1.46	1.41	1.15	1.55	0.83	1.60	0.91	1.37	1.62
	3	0.81	1.30	0.93	2.01	1.43	1.89	1.44	1.70	1.54	1.05
	4	0.75	1.13	0.85	0.66	0.88	0.87	0.58	1.47	1.40	1.07
	5	0.94	0.78	0.94	0.49	1.18	0.62	1.07	1.22	0.81	0.83