Universidade de Lisboa

Faculdade de Letras

Error analysis in automatic speech recognition and machine translation

NICOLAAS DIRK PETRUS LOOMANS

Thesis supervised by Prof. Sara Gonçalves Pedro Parente Mendes and co-supervised by Marina Sanchez, Ph. D.

2021

ABSTRACT

Automatic speech recognition and machine translation are well-known terms in the translation world nowadays. Systems that carry out these processes are taking over the work of humans more and more. Reasons for this are the speed at which the tasks are performed and their costs. However, the quality of these systems is debatable. They are not yet capable of delivering the same performance as human transcribers or translators. The lack of creativity, the ability to interpret texts and the sense of language is often cited as the reason why the performance of machines is not yet at the level of human translation or transcribing work. Despite this, there are companies that use these machines in their production pipelines. Unbabel, an online translation platform powered by artificial intelligence, is one of these companies. Through a combination of human translators and machines, Unbabel tries to provide its customers with a translation of good quality. This internship report was written with the aim of gaining an overview of the performance of these systems and the errors they produce. Based on this work, we try to get a picture of possible error patterns produced by both systems. The present work consists of an extensive analysis of errors produced by automatic speech recognition and machine translation systems after automatically transcribing and translating 10 English videos into Dutch. Different videos were deliberately chosen to see if there were significant differences in the error patterns between videos. The generated data and results from this work, aims at providing possible ways to improve the quality of the services already mentioned.

RESUMO

O reconhecimento automático de fala e a tradução automática são termos conhecidos no mundo da tradução, hoje em dia. Os sistemas que realizam esses processos estão a assumir cada vez mais o trabalho dos humanos. As razões para isso são a velocidade com que as tarefas são realizadas e os seus custos. No entanto, a qualidade desses sistemas é discutível. As máquinas ainda não são capazes de ter o mesmo desempenho dos transcritores ou tradutores humanos. A falta de criatividade, de capacidade de interpretar textos e de sensibilidade linguística são motivos frequentemente usados para justificar o facto de as máquinas ainda não estarem suficientemente desenvolvidas para terem um desempenho comparável com o trabalho de tradução ou transcrição humano. Mesmo assim, existem empresas que fazem uso dessas máquinas. A Unbabel, uma plataforma de tradução *online* baseada em inteligência artificial, é uma dessas empresas. Através de uma combinação de tradutores humanos e de máquinas, a Unbabel procura oferecer aos seus clientes traduções de boa qualidade. O presente relatório de estágio foi feito com o intuito de obter uma visão geral do desempenho desses sistemas e das falhas que cometem, propondo delinear uma imagem dos possíveis padrões de erro existentes nos mesmos. Para tal, fez-se uma análise extensa das falhas que os sistemas de reconhecimento automático de fala e de tradução automática cometeram, após a transcrição e a tradução automática de 10 vídeos. Foram deliberadamente escolhidos registos videográficos diversos, de modo a verificar possíveis diferenças nos padrões de erro. Através dos dados gerados e dos resultados obtidos, propõe-se encontrar uma forma de melhorar a qualidade dos serviços já mencionados.

Acknowledgements

First, I would like to thank my faculty supervisor, Sara Mendes. Her expertise, patience and professionality helped me throughout my academic journey and her advice and suggestions played an important role in executing the analysis and writing this thesis.

My thanks also go to my internship supervisor, Marina Sanchez. She always supported me during the internship and the period afterwards, and through her suggestions and recommendations she also contributed to the realization of this work.

I would like to thank Unbabel for offering me an internship and the aid that was necessary to write this thesis. This internship has contributed to my personal development, and I was able to learn new things. Its employees always assisted me when I needed help and answers on my questions.

Finally, I would like to thank my friends, family, and my girlfriend for the support they always gave me during good, but also psychologically complicated periods. They were the extra helping hand, that a person might need now and then.

Index

1. Introduction

"Break a leg!". For those who are not native English speakers, hearing this expression for the first time might be a bit of a shock. After all, you don't expect anyone to wish someone a broken leg in public. However, it is an expression used especially in theaters to wish someone good luck, but because it is mainly an expression that is common in England, a Portuguese and a Dutch person will most likely, in whatever possible context, not respond in a positive way, when you say "parte uma perna" or "breek een been".

Expressions like these are not easy to translate unless there is a literal equivalent for them. This applies to human translators, but certainly also to machine translation. If you enter "break a leg" as the input provided to a translation engine, a literal translation of the expression will be output, provided it concerns the language combinations English – Dutch and English – Portuguese. Nowadays, expressions could still be considered one of the Achilles' heels of translation engines.

For those who are unfamiliar with the term *machine translation* it can simply be said that it describes a process by which computer software translates text from one language to another without human intervention. Machine translation has become an important element in the translation process, as it is able to translate large amounts of text in a short time and, for many companies that use it, saves costs (Way A., 2018). Although machines are increasingly being used to translate texts, their quality is still variable, as is the case with expressions, for example.

A company that offers translation services and is concerned with ensuring their quality is Unbabel. Unbabel describes itself as a translation platform that helps companies to interact with customers in any language. A difference compared to translation agencies is the way in which they deliver their service. While many translation agencies only use human translation, Unbabel combines machine translation with human translation. For example, the text to be translated is put into a machine before being checked for quality by a human translator. By means of a multidisciplinary team, specialized in computational linguistics, Unbabel tries to provide customers with a translation of good quality.

In addition to machine translation, Unbabel also uses *Automatic Speech Recognition*, when translating videos, for example. The term *Automatic Speech Recognition* refers to an independent, machine-based process of decoding and transcribing oral speech and receives acoustic input from a speaker through a microphone, analyzing it using some pattern, model, or algorithm, to produce an output, usually in the form of a text (Lai, Karat, & Yankelovich,

2008). In translating videos at Unbabel, first an automatic system recognizes and transcribes the speech, then the text output by the automatic speech recognition module is put into a translation machine and then finally checked by a human translator.

However, the quality of the automatic speech recognition is not perfect and Unbabel is aware of this. Because Unbabel is so committed to the quality of the translation produced, it wants to verify the quality of both speech recognition and machine translation. In addition to the quality of these automatic systems, it is also interesting to find out to what extent Unbabel's speech recognition system affects the translation machine, as errors in speech recognition can potentially lead to errors in machine translation.

To achieve this, we translated videos with different characteristics after being transcribed by the *Automatic Speech Recognition* system. Research was conducted in relation to the types of errors made by the speech recognition system and the translation system and whether there is a link between the errors made by both. It is a very open type of research, in which we tried to find as much useful information as possible. This was done through an *annotation process*, a term that is further explained in another chapter. The mistakes made and any possible patterns are described in detail in the remainder of this work.

2. The company

Unbabel is a Portuguese start-up responsible for an artificial intelligence-driven translation platform that provides translation services through machine translation and human post-editing in real-time. The company was founded in 2013 by Vasco Pedro, João Graça, Sofia Pessanha, Bruno Silva and Hugo Silva and has its headquarters and branches in Lisbon, San Francisco, and New York. Unbabel has around 200 employees (2019) representing 27 countries and 17 languages. The company's mission is to create universal understanding and to achieve this, it combines machine translation with a community of 50,000 bilingual proofreaders for 29 languages and dozens of language combinations. In addition, it is mainly active in the customer service sector, which provides translations of emails, tickets, live chat, and FAQs (frequently asked questions). In addition, it offers video transcripts, translations, and subtitles.

Unbabel established a workflow in which the pattern of activity of the organization becomes clear. In general, Unbabel's translations are done as follows: the customer sends the source text (ST) to the company, it goes through a translation machine and is then divided into several small pieces of text that are checked and edited by a translator community. After this phase, the edited text goes through a *Quality Estimation* system. If the quality is not good enough, it is sent again to the editors. If it is good enough, it will be sent to the customer. There is a *Senior Editor* who reviews and evaluates the texts.

To indicate more clearly how Unbabel works, let's consider the pipelines presented below. These pipelines are regularly shown during presentations to customers, to help them understand how translation is performed at Unbabel. The translation pipeline below demonstrates how a text is converted into a delivered product (the final translation).

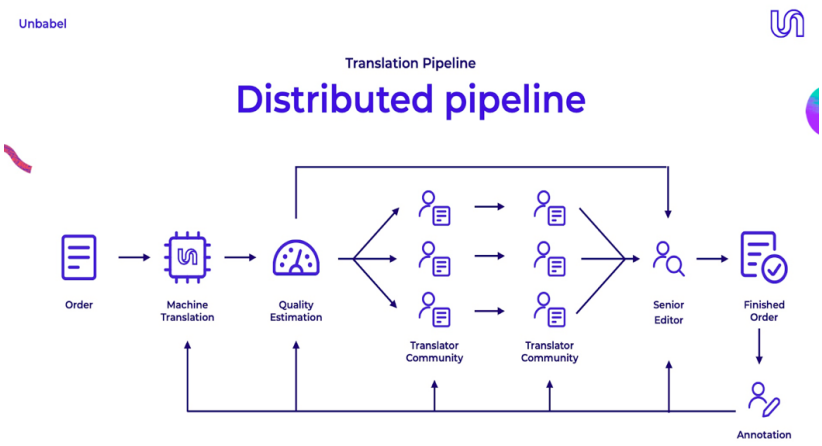

Figure 1: Unbabel's pipeline (Unbabel, 2019)

Before being translated, the source text is pre-processed. This means that the topic, genre, and difficulty of the text are defined, and glossary is prepared, when necessary. The glossary offers translation terms, depending on the content of each text and guaranteeing terminological consistency. Next to that, client-specific information is added. This could be, for example, general information about the company or instructions regarding the register and style of the text.

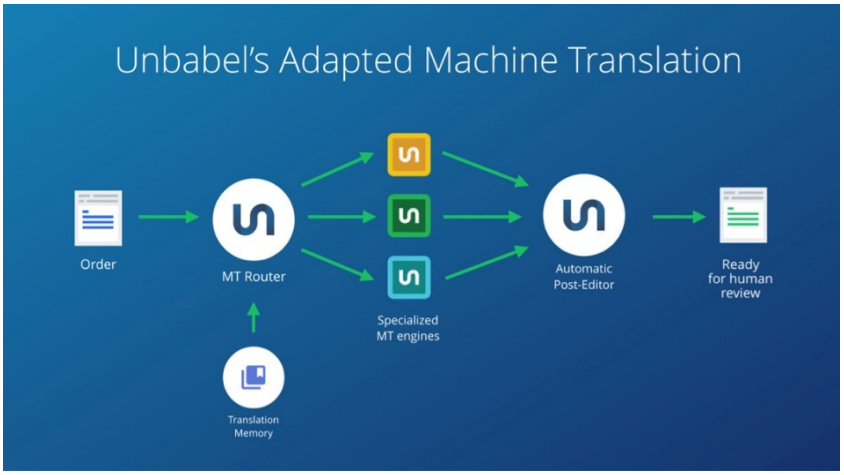

Figure 2: Automatic translation at Unbabel (Testa, 2018)

After the pre-processing of the text, the text is prepared for machine translation, starting with the *MT router*, shown in figure 2. This router chooses the best translation engines available at Unbabel, based on the content, area of expertise and client. After that, the text goes through the *APE (Automatic Post-Editor)*. With this tool, Unbabel improves machine translation quality

through automatic post-editing of errors identified by the *APE*. Then, the post-edited text is evaluated by the automatic evaluation system *OpenKiwi*, developed by Unbabel (Martins, 2019).

As shown in figure 1, the evaluated text is then sent to a community of editors (translators) or directly to a senior editor (also often called post-editor), depending on the quality of the text evaluated by *OpenKiwi*. The parts that lack quality, according to the system, are underlined.

After the pieces of text have been checked for quality by Open Kiwi, they are sent to the translator community, i.e., the regular translators or senior translators (also called post-editors or senior editors). These translators have access to the source text and the text translated by the translation engine. During this process, they also have access to auxiliary tools (NLP tools), including *Smartcheck*, translation memories, glossaries, and a spellchecker.

Translation memories are databases of fragments of texts that consist of valid translations that were already used before. This database can also provide contexts that assist on, for example, gender and number issues. This is illustrated in the image below, that was used in a thesis of a former student that worked at Unbabel.

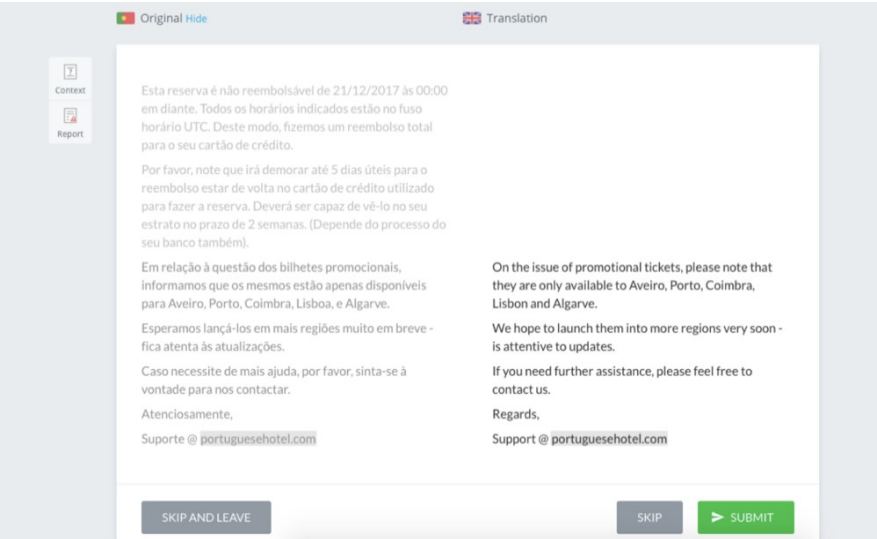

Figure 3: Example translation memory (Taysk, 2019).

Smartcheck is a static analysis tool that detects errors. Unbabel, together with other researchers, has succeeded in further developing this tool to adapt it to its own service. This means that the tool helps the Unbabel community of translators with proofreading by providing

them alerts and suggestions, related to spelling, register, lexical coherence (subject-verb agreement, correspondence between pronouns, gender, etc.) and other rules related to customer requirements. This contributes to the quality of Unbabel's translations, speeds up the translation process and helps translators not only by pointing out possible errors, but also by providing helpful hints to correct them.

The tool signals errors or hints by underlining the words in green or red. In the first case, translators can decide whether to make any changes or not in the translation and the error is not considered as one that should really be changed. Errors that are underlined in red, the second case, are considered as critical errors and leave the translator almost no other option than correcting them before the translation is delivered. When selecting an underlined word, *Smartcheck* provides a list of suggested words, which can be accepted or ignored with a simple click.

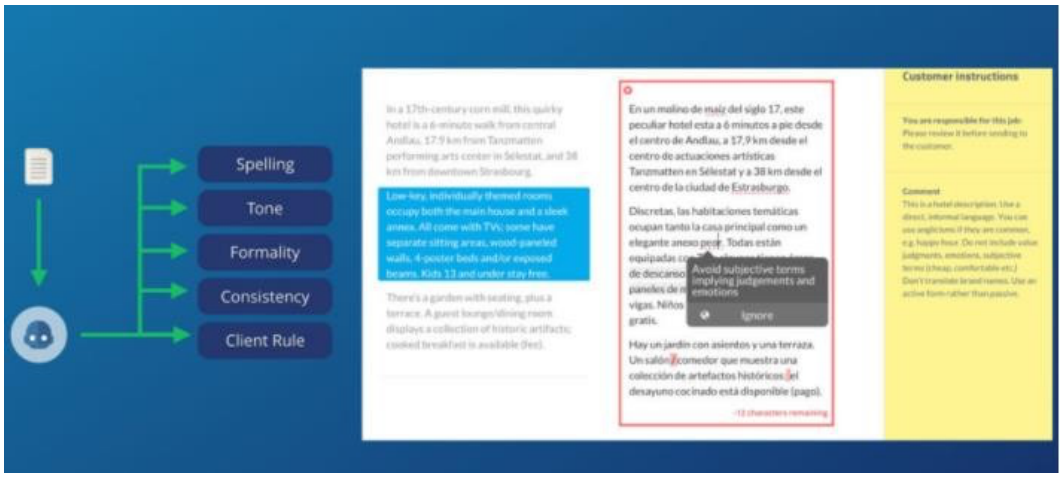

Figure 4: Smartcheck example (Testa, 2018)

Other tools to be mentioned are the *Dependency parser* and the *Turbo parser*, as they have an important function in improving the quality of the texts by *dependency parsing*. *Dependency parsing* is the task of recognizing a sentence and assigning it a syntactic structure. The *Dependency Parser* is a useful tool for translation, as it can emphasize syntactic relationships between words and sentences and, by means of *part-of-speech tagging*, automatically assign a lexical category to each word (Testa, 2018). It solves syntactical and ambiguity issues, depending on the relation between constituents and the meaning of a constituent depending on the part-of-speech.

Unbabel uses a similar tool called the *turbo parser*.

> *We present fast, accurate, direct nonprojective dependency parsers with third order features. Our approach uses AD3, an accelerated dual decomposition algorithm which we extend to handle specialized head automata and sequential head bigram models. Experiments in fourteen languages yield parsing speeds competitive to projective parsers, with state-of the-art accuracies for the largest datasets (English, Czech, and German)* (Martins, Almeida, & Smith, 2013, p. 1).

This tool helps to analyze data and provides finer-grained information to the *Smartcheck* thus allowing it to improve the suggestions provided. Information is provided at word level and takes into account its part-of-speech, and specific features, such as number, gender, person, mood, tense, or verb form.

The pipeline previously presented, however, only applies to three of the four Unbabel products: Tickets, FAQs and Chat. A slightly different pipeline applies to videos, and is presented below:

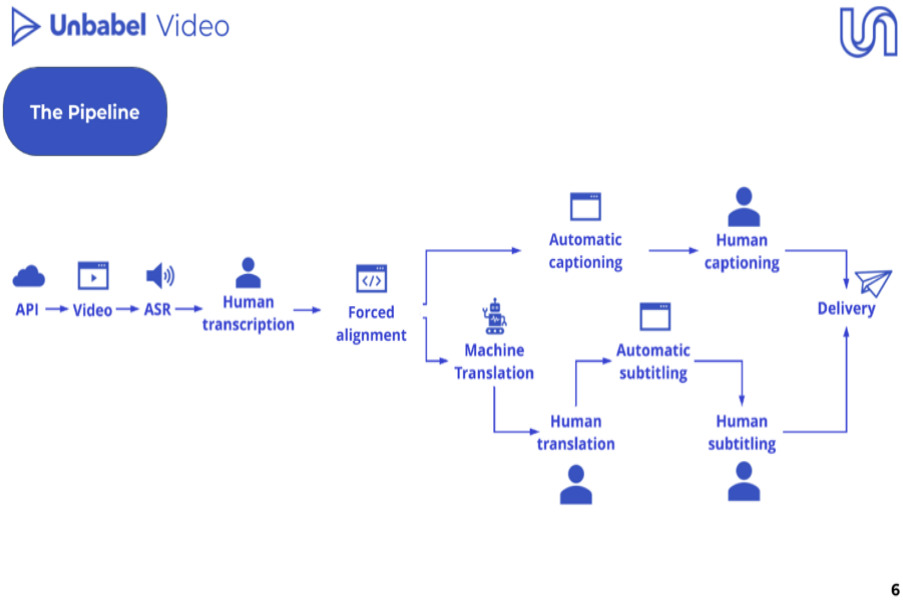

Figure 5: Pipeline for video (Unbabel, 2019)

For videos, automatic speech recognition is used before the text is sent for transcribing. This is one of the differences compared to the pipeline for other products. Then, depending on the quality of the transcription, it is either put in a translation machine and then checked by human translators and subtitlers, or immediately sent to a subtitler.

As previously mentioned, Unbabel is very concerned about the quality of the translations (Martins, Almeida, & Smith, 2013). To ensure their quality, it makes use of the earlier described community of editors and post-editors, whose performance is regularly assessed by the "Evaluate" system.

Next to that, there are processes that are carried out by professional linguists called *annotators* (not to be confused with editors or post-editors/proofreaders) who assess the quality of the proofreaders and of the final translation delivered to the customer. This is done through the annotation process, which is made possible by an annotation tool. In the image below, a graphical representation of the annotation tool is shown.

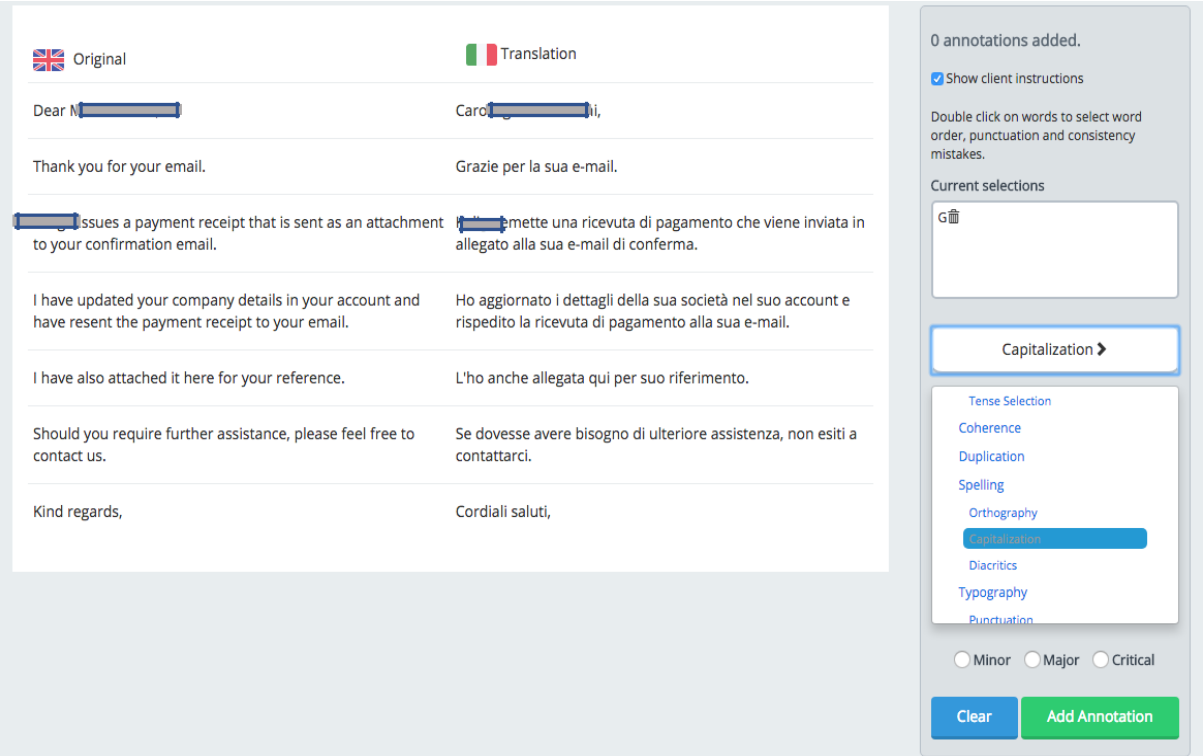

Figure 6: Example of Unbabel's annotation tool (Testa, 2018)

Annotators identify errors in the translated text. On the left part of the window, the source and target text are shown (in this case a translation of English into Italian). The annotators can select the text that they want to evaluate. Then on the right, the annotator can choose the type of error that is made. As shown in the image, there are several types of errors that the annotator can choose from, such as "Capitalization", but it can also define the severity of the error (*minor*, *major* or *critical*). In this annotation process, Unbabel uses an error typology of the MQM model, which is a framework for describing quality metrics used to assess quality and identify

specific issues in translated texts. More information about this model is described in the "Literature review".

Annotators are professional translators with at least five years of experience and evaluate the performance of the editors by analyzing everything from the original text to the delivered text; they comment on the last edited text and rate it on a scale of 1 to 5 (1 for the worst performance and 5 for the best).

Editors, post-editors, and annotators can also use the report feature on their platforms to register technical issues in each of these tools. In fact, all these tools are very useful for the company's processes, but because Unbabel is a startup, it regularly implements new features and tools, which are updated and improved with user feedback.

To facilitate the process for editors, post editors and annotators alike, Unbabel defined a set of guidelines. Editors will find client instructions and register and style recommendations in the *Translation Guidelines* while, for example, annotators will find the error typology used at Unbabel in the *Annotation Guidelines*. Like all the other tools and processes described, the guidelines play an important role to ensure the quality of the translations delivered by Unbabel. The *Annotation Guidelines* are particularly important, and they are one of the aspects discussed in detail in this thesis. Being so, they can be found in the appendices.

3. Literature review

This section describes models and relevant sources for the research. The first tools that should be mentioned are the ones that are considered in detail in this chapter, ASR and MT systems. It is relevant to define and describe these tools because their characteristics have an impact on their performance.

3.1 Automatic Speech Recognition

A study on ASR developed at the University of TAHRI Mohamed, Algeria, gives an overview of the main definitions of ASR and provides a summary of relevant research on speech processing in the last few years (Benk, Dennai, & Elmir, 2019).

Automatic speech recognition, also called *speech recognition*, can be defined as graphical representations of frequencies emitted as a function of time and allows a machine to understand the user's speech and convert it into words through a computer program.
According to researchers, such as the researchers at the University of TAHRI Mohamed, but also at, for example, the Tata Institute of Fundamental Research (Samudravijaya, 2008) speech recognition systems can be categorized in different groups, the most important being those that focus on the nature of the utterance, size of vocabulary and number of speakers.

In most studies, it is shown that utterances can be of four types: *isolated words, connected words, continuous speech, and spontaneous speech*. And ASR systems are often trained to deal with one of this type of utterance:

- *Isolated Word Recognition systems*: a type of system in which a user is required to pronounce words with a clear pause between them.
- *Connected Word Recognition systems*: a type of system that recognizes words from a small set and is comparable with the *Isolated Word Recognition* system, but it allows separate words to be pronounced together with no need for a pause between them.
- *Continuous speech recognition systems*: A type of system that recognizes sentences spoken continuously, while the computer selects the content that is spoken.
- *Spontaneous speech recognition systems*: A type of system that recognizes natural-sounding and not rehearsed speech; these handle speech disfluencies such as *ah* and *um*, or grammatical errors in conversational speech.

The size of the vocabulary is considered important because it determines the accuracy of the speech recognition system. Some systems only consider a few words, while others deal with many words. Regarding the vocabulary size, researchers describe the systems as follows:

- Small vocabulary systems - systems that use around tens of words.
- Medium vocabulary systems - systems that use hundreds of words
- Large vocabulary systems - systems that use thousands of words
- Very-large vocabulary systems - systems that use tens of thousands of words

As for the speakers, there is a distinction between systems that depend and those that do not depend on the speakers. A system is *speaker independent* if it can recognize speech of any and every speaker and has learnt the characteristics of many speakers. Within this category, *speaker adaptive* systems can be further distinguished. These systems can adapt to the voice of a new speaker, considered that enough speech is provided for training it. A *speaker dependent* system is not able to recognize new speakers well, meaning that they are dependent on the data that is used for training.

Research in ASR systems has been around for a long time and several developments have occurred over time. To understand ASR systems even better, it might be important to look at its history. The section below provides a brief description of the history of ASR systems.

3.1.1 History of ASR systems

3.1.1.1 Early interest in speech processing

Interest in speech processing started to become visible a few centuries ago. In fact, at the time the interest focused on developing speech machines. In the 2nd half of the 18th century, 1773 to be more precise, Danish scientist Christian Gottlieb Kratzenstein built models of the human vocal tract that were able to produce five vowels (Kratzenstein, 1782).

About 20 years later, in 1791, an acoustic-mechanical speech machine was introduced by Wolfgang von Kempelen (Dudley & Tarnozcy, 1950). The advantage of this machine was that, due to the specific model of the human vocal tract, it was able to produce single sounds, but also some combinations of sounds. The machine had a pressure chamber mimicking the lungs, a vibrating reed acting as vocal cords and a leather tube for vocal tract action. By manipulating the shape of this tube, it could produce different combinations of sounds. At the time, von Kempelen was faced with negative publicity and was not taken seriously, because some of his inventions were proved fraudulent. Nevertheless, his machine ensured new theories regarding human vocals (Svendsen, 2003).

Based on this machine, another version was created in the mid-19th century by Charles Wheatstone, which, compared to von Kempelen's machine, could produce vowels and most

consonants, and even some full words. In 1881, Alexander Graham Bell constructed a machine that is very similar to Wheatstone's machine (Huang & Baker, 2014).

Another notable invention is the VODER, a speech synthesizer developed by research physicist Homer Dudley in the 1930s. This synthesizer was invented as a result of research into techniques for telephone voice encryption at Bell Laboratories (Dudley, Riesz, & Watkins, 1939) and is an almost identical machine to Wheatstone's, albeit electrically, not mechanically. The image below roughly shows how the model worked.

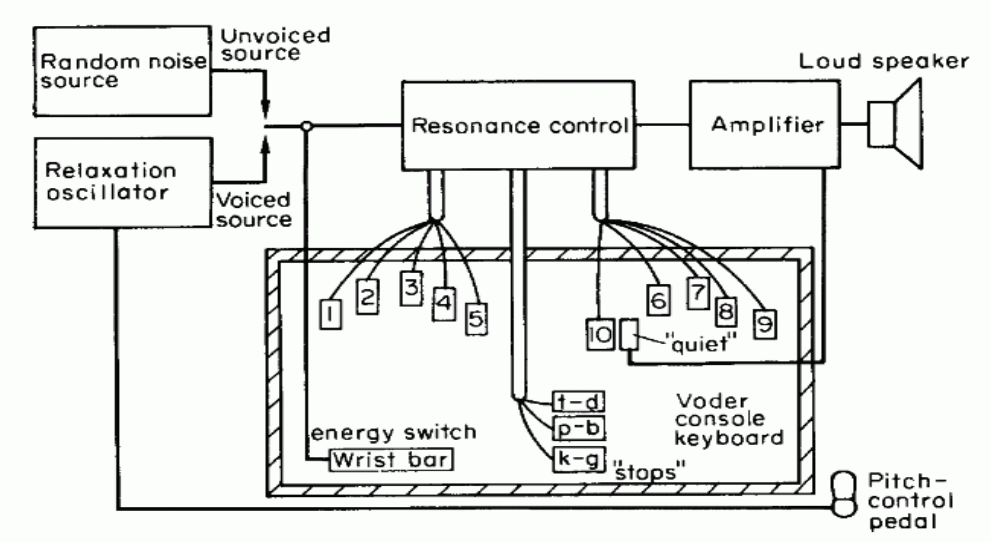

Figure 7: Model of Dudley Homer's VODER (Dudley, Riesz, & Watkins, 1939)

The operator of the VODER could select an output of a so-called *relaxation oscillator*, a "nonlinear electronic oscillator circuit that produces a nonsinusoidal repetitive output signal, such as a triangle wave or square wave" (Morris, 1992) or a random noise as driving signal by using a wrist bar (see image). Next to that, there was a pitch-control pedal to control the oscillator frequency. This driving signal was then passed through ten bandpass filters (represented by the ten numbers in the image) whose output levels were controlled by the operator's fingers. These filters were used to change the power distribution of the source signal across a frequency range, being able to determine the characteristics of the sound at the loudspeaker. This means that to synthesize a sentence, the operator needed to learn how to control the VODER to produce the right sounds of the sentence. The VODER is considered an important milestone in the evolution of speaking machines, especially because it was even shown at the World Fair in New York City in 1939 (Williams, 1940).

3.1.1.2 First recognizable ASR systems

The section above showed that interest in speech processing came into being a long time ago. However, this early interest was not on recognizing speech, but on creating a speaking machine (Juang & Rabiner, 2005).

What can be considered as the first *Automatic Speech Recognizer* was a system called *Audrey*, created by researchers at Bell Laboratories in 1952 (Davis, Biddulph, & Balashek, 1952). This was a system that was able to recognize numbers spoken by a single voice. Back then, computer systems were not flexible and were very expensive, with limited memory and computing speed. Still, Audrey was able to recognize the sound of a spoken digit (zero to nine) with over 90% accuracy, provided it was spoken by its own developer HK Davis. With other speakers, it had a 70-80% accuracy rate. With unfamiliar voices, the performance of the system was visibly worse. According to Charlie Bahr, employee at Bell Labs Information Analytics at the time, "This was an amazing achievement for the time, but the system required a room full of electronics, with specialized circuitry to recognize each digit" (Moskvitch, 2017).

Another system that could be considered worth mentioning, is IBM's *Shoebox* machine, created by William C. Dersch in 1961. IBM states that it was "a forerunner of today's recognition systems" (IBM, 1961). It was a device that was operated by speaking into a microphone, which converted sounds into electrical impulses, and was able to recognize and respond to 16 spoken words, and to the digits 0 to 9. It was also a machine that was able to solve arithmetic problems. When words such as *plus*, *minus* and *total* were spoken, Shoebox calculated and printed answers to simple calculations. In 1962, Dersch himself demonstrated this machine at the World Fair in Seattle.

3.1.1.3 Systems in the 70s

The progress that was made in previous years continued in the 1970s, beginning with the five-year program called *Speech Understanding Research* of the U.S. Department of Defense's ARPA in 1971. 3 million dollars were spent in this project leading to several Speech Understanding Research groups, creating new ASR systems, and was the largest speech recognition project ever (Huang & Baker, 2014). One of the outcomes of this large-scale project was the *Harpy* system, developed by the Carnegie Mellon University in 1976. It was able to understand 1011 words, approximately the vocabulary of a three-year old child (Juang & Rabiner, 2005). Harpy was considered a system that was significant because it introduced a search approach called *graph search* to "prove the finite state network of possible sentences" (Waibel & Lee, 1990). In this graph search, "the speech recognition language was represented

as a connected network, derived from lexical representations of words, with syntactical production rules and word boundary rules" (Itakura, 1975). Harpy was one of the first systems that made use of *finite state networks* and until the early 90s there have been virtually no systems that optimized this network (Mohri, 1997). Several systems emerged from ARPA's project, but the Harpy system is considered the most noteworthy. This might have to do with the fact that the systems did not meet ARPA's goals of the project (Klatt, 1977).

In addition to ARPA-funded work, research from IBM and AT&T Bell laboratories also came into being in the same decade. Both companies endeavored to examine the applicability of automatic speech recognition systems for commercial applications. However, they had different goals and focus. IBM tried to create a *voice activated typewriter* (VAT), whose main function was to convert a spoken sentence into a sequence of letters and words that could be shown on a display or typed on paper (Jelinek, Bahl, & Mercer, 1975). This system, called Tangora, was a speaker-dependent system, as it had to be trained by each different user. In terms of its development, it focused on the size of the vocabulary to be recognized and what was called a *language model*, i.e. a set of statistical grammar rules. For this language model, they used an *n-gram model*, which defined the probability of sequences of words (*n)*. This n-gram model later was used regularly in multiple systems that focused on the size of the vocabulary.

The goal of AT&T's research was to provide automated telecommunication services to the public, such as voice dialing (Juang & Rabiner, 2005). The systems that the company wanted to develop were supposed to work for a large population of speakers and they needed to be *speaker-independent*, meaning that it would not be necessary to train the systems with individual speakers. Applications at the time, such as voice dialing, usually were trained by short utterances and limited vocabulary, consisting of only a few words. Because of this, Bell Laboratories wanted to focus on what was generally called an *acoustic model*. This model roughly consisted in a spectral representation of sounds or words rather than a representation of the grammar or syntax, such as the representation used in the language model of IBM's Tangora. What also might be noteworthy in AT&T's approach was the concept of *key spotting*. Key spotting aimed at detecting a keyword or phrase in a longer utterance that was not semantically significant to other words in that utterance. This was to accommodate speakers that preferred to use natural sentences. For example, a telephone caller, when requesting services, just needed to say the word "credit card" rather than using a naturally spoken sentence to make the system understand that speaker wanted to make a credit card call.

IBM's approach, AT&T's approach and the ARPA-funded projects had a significant influence in the evolution of speech technology. However, despite the progress made in the 70s, these were not considered a period of great success (Juang & Rabiner, 2005). Despite the developments made, most systems were only able to recognize a small number of words.

3.1.1.4 Systems in the 80's and 90's

The 80s and 90s were a period in which ASR systems developed drastically in terms of vocabulary. Thanks to these new developments, the systems were able to recognize thousands of words instead of hundreds. The first model or approach that contributed to this trend is *The Hidden Markov Model* (Pinola, 2011). What needs to be noted is that the emphasis of this model was different. "Markov's approach represented a significant change from simple pattern recognition methods, based on templates and spectral distance measure, to a statistical method for speed processing (Rabiner, 1989)". Next to that, the model considered the variability of speech signals and the structure of a spoken language. When people say the same word, it is possible that the acoustic signals are not quite the same, even though the linguistic structure is the same in terms of, for example, grammar, syntax, and pronunciation. This happens in, for example, dialects. By trying to measure the probability of unknown sounds being words, considering the variability of speech signals and the structure of the spoken language, it was able to recognize much more words than other systems in previous years. Afterwards, the model turned out to be very successful and several models were developed based on this one for decades (Pinola, 2011).

Due to the success of the Markov model, besides the model itself, there were few noteworthy systems or approaches in the 80s. In 1982 and 1984, respectively, *Dragon Systems* (playing an important role later) created by two doctors Jim and Janet Baker, and *SpeechWorks*, which at the time was a leading provider of over-the-phone automated speech recognition, were founded (Huang & Baker, 2014). What should be mentioned, however, is that in 1987 an effort was made to develop (or further develop) something that characterizes the 90s in the field of speech recognition: software tools. These tools were mainly intended for business and specialized industry, but they even reached the general public. In 1987, by means of 'Worlds of Wonder's Julie doll', children were able to train a doll to respond to their voice (Swamy & Ramakrishnan, 2013). The problem however was that the programs at the time took discrete dictation, causing it to be necessary to pause after each word (Pinola, 2011).

Nevertheless, in the 90s great progress was made as systems became more sophisticated, making the dictation software more advanced. What is important in this part too, is that in the

90s more speech recognition systems became accessible to the general public. In 1990, Dragon Systems Inc. launched the first consumer speech recognition product, *Dragon Dictate*, with a price of 9000 dollars. Seven years later, the same company also launched *Dragon NaturallySpeaking*. This system was, logically, more advanced and was able to recognize 100 words per minute, which was a significant improvement compared to systems in the 80s, as it was no longer needed to pause between words for the computer to understand what was being said (Huang & Baker, 2014). Another discovery from the 90s is the first *voice portal*, the *VAL* from BellSouth in 1996. *VAL* was a dial-in interactive voice recognition system that gave information based on what was said. Charles Schwab's program *Voice Broker* developed in the same year is noteworthy too. The program allowed 360 customers at the same time to call in and get information about stocks and options and had an accuracy of 95% (Juang & Rabiner, 2005).

3.1.1.5 2000's till now

During the early 2000s, the speech recognition area was still dominated by *Hidden Markov Models* in combination with *Artificial Neural Networks* (Bourlard & Morgan, 1994). An example of a development worth mentioning are speech recognition programs *EARS* (Effective Affordable Reusable Speech-to Text) and *GALE* (Global Autonomous Language Exploitation), both sponsored by DARPA. The EARS program was led by four participants: IBM, BBN Technologies, Cambridge University, and a team composed of ICSI (International Computer Science Institute), SRI (Stanford Research Institute) and University of Washington. Its goal was to "significantly advance the state-of-the-art while tackling the hardest speech recognition challenges including the transcription of broadcast news and telephone conversations" (University of Cambridge, 2002). *EARS* financed *Switchboard*, a large multispeaker corpus of conversational speech and text, containing about 2500 conversations by 500 speakers from around the US (Godfrey, Hollman, & McDaniel, 2002). The *GALE* program tended to develop and apply computer software technologies to absorb, translate, analyze, and interpret huge amounts of speech and text in multiple languages. This project was active for 2 years in which the participants were able to develop speech recognition, translation, and information delivery systems in Chinese and Arabic (Cohen, 2008).

The subsequent period was characterized by the increasing use of neural networks. An example of an approach that was based on neural networks was the *deep learning* approach, an approach that also started to be used for machine translation. *Deep learning* defines a subset of machine learning, which is essentially a neural network with three or more layers, attempting

to simulate the behavior of the human brain allowing it to learn from large amounts of data (IBM, 2020). Neural networks started to be explored in the 80's and 90's, although at the time they were not yet able to dislodge the speech recognition systems based on the *Hidden Markov Model*, as they faced some obstacles in combination with the lack of training data at (Deng, Hassanein, & Elmasry, 1994). From the early 2010s, researchers started to overcome these obstacles and lack of training data, and systems based on neural networks started to become dominant in the speech recognition area. Deep learning decreased word error rate by 30% and was quickly adopted across the field (Markoff, 2012). Well-known examples of systems that are based on neural networks are *Google Voice* and Apple's *Siri*.

3.2 Machine translation

Machine translation refers to the attempt to automate the process of translating natural language utterances from one language to another (Arnold, 1994). Machine translation is a *Natural Language Processing* system which uses a bilingual data set to build language and phrase models used to translated text. When talking about machine translation, there is no human involved and the text is exclusively processed by computers. Therefore, machine translation should not be confused with *computer-assisted translation*, as the translations for *computer-assisted translation* are made by humans (Costales, 2009).

Due to globalization, companies nowadays are communicating more than ever with each other on an international level. This communication still causes a lot of problems and to solve these, more and more companies are using machine translation, mainly because it is fast and cheap (Peng, 2018). It is also known, however, that machine translation outputs are not perfect yet and bring some issues (Stankevičiūtė & Kasperaviciene, 2017). Examples of these issues, for example, are the inability to account for local phrases due to lack of context, difficulty to accurately translate nuances, slang and other culturally relevant phrases, and the possibility for brand damage due to a lack of cultural awareness and cohesiveness.

Regarding the quality of machine translation, there are still several question marks:

> *Results obtained with MT processes are variable and depend on different factors, such as the genre and domain of the source text, the aim of the text, and the syntax and the lexicon. Most of the time, the generated text is a "raw" translation: its quality is poor (Testa, 2018, p. 5).*

As stated above, the quality of machine translation depends on different factors and the fact that it is so variable, results in the fact that human translation is still preferred by parties that

benefit from good quality translations (Rojo, 2018). Nevertheless, translation machines are increasingly used. To get a better idea of the developments that machine translation went through, a brief description of its history is given in the next section.

3.2.1 History of machine translation

For the first signs of machine translation, we have to go back in history. Mel'čuk and Ravič (Melčuk & Ravič, 1967) talk in their bibliography about the earliest known attempted mechanical translation system. They claim that the system, which seemed to be a prototype mechanical translating typewriter, was reported in an Estonian newspaper called *Vaba Maa* on the 24th of February in 1924.

The first systems for which detailed information is available are of French and Russian origin. George Artsruni, a French engineer, developed a system based on a paper tape (Corbé, 1960), which was publicly demonstrated at the Paris Universal Expo in 1937. It was not a complete translation machine, but should rather be considered as a mechanical bilingual dictionary. Another system of which there is ample evidence is that of Petr Petrovič Trojanskij (Hutchins & Lovtskii, 2000), who proposed a three-part translation process, which was the first to came up with the use of post-editors (and pre-editors). However, his proposal was never actually built.

The first attempts to achieve full automated translation began in 1949, after the Second World War. In this year, Warren Weaver of the Rockefeller Foundation (philanthropic institution in New York) began a correspondence with Norbert Wiener, a professor at the Massachusetts Institute of Technology, in which they tried to examine the possibility of using computers to translate (Weaver, 1949). The ideas discussed in this correspondence were expanded in a memorandum published by Warren Weaver.

This memorandum was inspired by the success of *cryptography* during World War II[1] and Weaver stated that translation of human languages could be conceived as a problem of *cryptography*. It was the first publication of the 20th century in American and Western Europe that was known to indicate the possibility of using computers to make translations, and, also important, that cryptography methods might be useful for machine translation. Weaver proposed methods to solve ambiguity, which was a well-known linguistic issue in natural language texts. He also acknowledged that basic machine translation might be useful for the

[1] Cryptography is the practice of techniques for secure communication in the presence of third parties. It was used extensively in World War II to protect information and communication (Budiansky, 2000).

translation of technical and scientific documents but would probably lack quality in the translation of literary texts. The memorandum is widely recognized as the starting point of machine translation in the mid-20th century (Schwartz, 2016).

After this came a century referred to as "the century of optimism" (Hutchins, 2014). This optimism came after a survey that was published by Yehoshua Bar-Hillel, a full-time machine translation researcher, in 1951 (Bar-Hillel, 1951). In this survey he discussed the state of the art in machine translation and foresaw important problems that would be encountered in the coming years. To deal with these problems, he proposed possible human-machine partnerships, wherein humans could serve as pre-editors or post-editors to MT systems. Bar-Hillel brought this survey and Weaver's memorandum to a conference organized by himself in which ideas and perspectives on machine translation were presented by him and other participants. Even though these ideas were considered insightful but unrealistic (Hutchins, 2014), this was the moment after which the optimism in relation to the development of machine translation started.

This optimism was further reinforced by the first public demonstration of machine translation by researchers at IBM and Georgetown in 1954 (Dostert, 1955). This was a demo of a system that used a small vocabulary to translate a fixed set of sentences from Russian to English, that was widely recognized as a "resounding success" in the press (Hutchins, 1999). Despite of its limitations, this demo created a lot of enthusiasm, causing an increase in research in automatic translation in the United States, but also in Western Europe in the following years.

In the early 1960s, however, this positivism turned into negativism. It was Bar-Hillel himself, who was so positive earlier, who expressed this negativity.

> *During the first years of the research in MT, a considerable amount of progress was made which sufficed to convince many people, who originally were highly skeptical that MT was not just a wild idea. It did more than that. It created among many of the workers actively engaged in this field the strong feeling that a working system is just around the corner. Though it is understandable that such an illusion should have been formed at the time, it was an illusion* (Bar-Hillel Y. , 1960, p. 100).

This excerpt is taken from a second survey Bar-Hillel conducted. In the remainder of this survey, he states that people, including himself, had been too skeptical, and, as shown in the excerpt, that the high expectations in relation to machine translation were an illusion. In this survey, Bar-Hillel also stated that fully automatic high-quality machine translation (FAHQT) was unachievable. The report aroused attention but did not lead to changes in research direction or techniques among other MT researchers (Hutchins, 1999).

In part because of this, the Automatic Language Processing Advisory Committee (ALPAC) was established by the US National Academy of Sciences to review the results of US research on machine translation in 1964 (ALPAC, 1966). The committee's report was extremely negative: the costs of developing translation systems would far outweigh the benefits, and following what Bar-Hillel had already stated, it was an illusion that machines could deliver good quality translations. Bar-Hillel's idea to make use of pre and/or post-editor partnerships was not accepted, probably because another project of researchers at Georgetown in 1962 had already shown that a translator needed less time for translating a text from scratch than to correct machine translated text (ALPAC, 1966).

Despite the consequences of these negative publications, research did not stop completely. In the 1970s, a few universities continued to develop research in machine translation, the universities of Grenoble, Heidelberg, Saarbrücken, Texas, Montreal, and Hong Kong, in particular (Hutchins, 2001). In addition to universities, there were also other organizations that produced machine translation systems, such as LOGOS.

At the end of the 1970s, interest in machine translation was on an upward trend again. In 1978 a handbook by Bruderer appeared: an 800-page inventory of translation machines under development and of devices that could be of use to the translator (automatic dictionaries, etc.) (Bruderer, 1978). In the same year, a permanent working team was also set up by the European Economic Community to develop a translation system called *Eurotra* for the languages of the EEC (Campbell & Cuena, 1989). One of the main reasons, however, for the return of a slight positivism is that some promising translation machines, such as *SYSTRAN* and *TAUM*, were developed.

SYSTRAN was developed in an industrial environment by Peter Toma, who had a strong connection with Georgetown University. The system had different versions. The Russian-English version was used by the United States Air Force (ASAF) during the Cold War. The English to French version, initially made for the Canadian market, was purchased by the EEC in 1975. Later the French to English and the English to Italian versions were also purchased. The purchase of these systems did not mean that they could be used immediately. Numerous improvements had to be made, so that, for example, “nous avions” was no longer translated into “we airplanes”. Only from 1981 on, after several years of development, it sometimes made sense to have SYSTRAN make a translation, although unfortunately, in some cases, the translation was so bad that it had to be thrown away completely (Neijt & Hoekstra, 1986).

The TAUM group was a research group at the university of Montreal that mainly worked in machine translation from 1968 till 1980. In the translation system developed by the

group itself, four people created *TAUM-Météo*, a program for translating weather reports from English into French, in two years. At the time, this system was considered a very successful translation system: more than 80% of the sentences were correctly translated by the machine. As to the other 20% of the sentences, the machine did not know what to do. Those sentences were automatically forwarded to human translators. The results were promising, but critics said to keep in mind that the system translated a very limited type of text. *TAUM-Météo* could not therefore be seen as a guarantee for success in translating text types with more variation and language aspects, such as expressions for example, that are difficult to translate.

During the 1980s MT advanced rapidly on many fronts. Many new operational systems appeared, the commercial market for MT systems of all kinds expanded, and MT research diversified in many directions (Hutchins, 2001). One example of those commercial systems was LOGOS. LOGOS already translated aircraft manuals during the 70s from English to Vietnamese, but a more eye-catching project was a German-English system for telecommunication manuals that appeared on the market in 1982. This system was later bought by the Commission of the European Communities.

But Japan was where most of the industrial work in machine translation was done in the 1980s. Most computer companies had developed software for computer-aided translation, mainly for the language pairs Japanese to English and English to Japanese, but there was also a lot of demand for Korean and Chinese. Examples of such systems are *AS-TRANSAC* (Toshiba), *MELTRAN* (Mitsubishi) and *ATLAS* (Fujitsu).

However, the most advanced available commercial system in the 1980s was the *METAL* system, which was realeased in 1988, developed by researchers at the University of Texas. The system was initially developed to translate documents in data processing and telecommunications from German to English. Later this system was followed by other systems, covering languages such as Dutch, Spanish, French, but also English and German.

The 90s were characterized by the introduction of a new approach (Hutchins, 2001). While the most common approach until the end of the 1980s was the *rule-based approach* based on linguistic rules, in the 90s this changed to the so-called *corpus-based approach*, in which rules were deduced from corpora. This approach was first introduced by a research group at IBM, who developed a system called Candide, based purely on statistical methods. Statistical methods were common in the beginning of the 60s, but the results, in general, were disappointing. Nevertheless, to the surprise of many companies that used the *rule-based approach* (based on linguistic rules), IBM was able to deliver a system with "acceptable results" (Hutchins, 2001).

Another example of a corpus-based approach is an approach known as an *example-based approach*. Such an approach was first proposed in 1984 by Makoto Nagao, but the experiments did not start until the end of the 80s. The example-based approach is built on extracting and selecting from a databank equivalent phrases or word groups, which were adapted by statistical methods or by more traditional rule-based methods. This means that example-based approaches can be approaches either based on rules or on statistics, but the main feature is that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents (Hutchins, 1999).

In the 90s there was also an increase in the demand of machine translation due to the fact that internet started to be more used (Kenny, 2018). Because of this, software that was specialized in emails and webpages started to develop. According to Hutchins (2016), an example of this kind of software was SYSTRAN. Also, since the beginning of the 90s, several other machine translation developers started providing machine translation services online, such as *Babel fish*, which was launched as a subdomain of the *AltaVista* search engine, but also *Reverso* and *PARS*. The quality of online machine translation services was poor, but it was enough to get the general meaning of the text.

The 2000's (especially the first ones), due to the large number of translation services available online, can be seen as the years in which *statistical-based systems* were popular (Testa, 2018). The first ideas of statistical translation machines were already introduced by Warren Weaver and were re-introduced in the late 80s and early 90s (IBM's Candide), but there was a renewed interest on them in the early 2000's. At that moment, it was by far the most widely studied machine translation method (Brown, et al., 1990). Statistical Machine Translation (SMT) systems were getting widespread due to their good performance, and were considered the most advanced and efficient form of machine translation until the launch of Neural Machine Translation (NMT) systems developed between 2015/16 and quickly adopted by companies like Google, SYSTRAN and Microsoft (Koehn, 2016).

The most recent type of translation and the successor to the statistical translation machines is the *Neural Machine Translation* (NMT), mentioned above. NMT departs from phrase based statistical approaches that use separately engineered subcomponents ordering (Wolk & Marasek, 2015) and it takes inspiration from the neural system of the human brain. According to Bentivogli, Bisazza, Cettolo, & Federico (2016), NMT became the new state-of-art, especially if it comes to languages pairs involving rich morphology prediction. "NMT output contains less morphology errors, less lexical errors, and substantially less word order errors" (Koehn & Knowles, 2017). On the other hand, is also mentioned that "NMT systems

have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency". However, as NMT is still a recent translation method, it is difficult to generate any significant findings on it. According to (Castilho, Moorkens, Gaspari, Popovic, & Toral, 2019), we are still at the beginnings of the development of NMT systems, so it is still necessary to make more in-depth research with larger samples, involving more pairs and considering different levels of experience, such as the use of results of this type of system in post-editing or pre-editing processes.

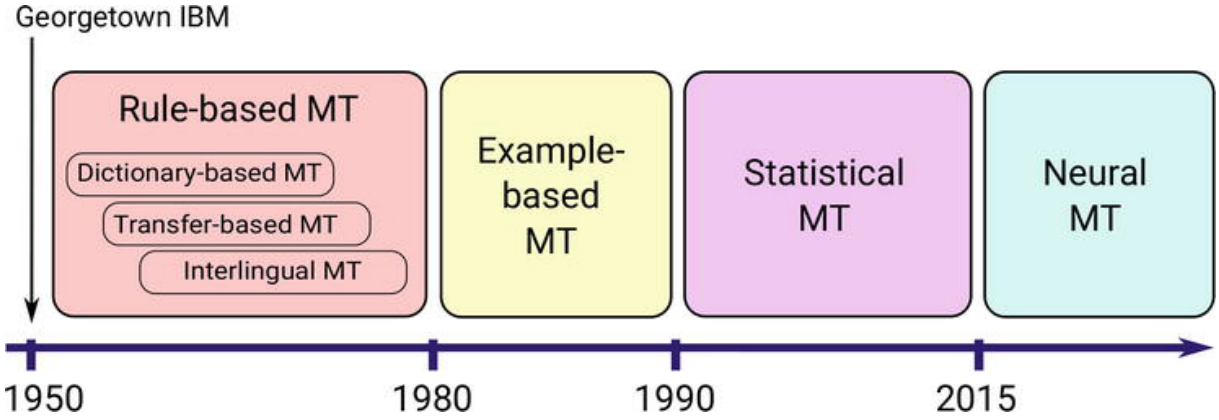

Figure 8: Chronological Timeline Machine Translation (Maučec & Donaj, 2019)

The image above gives a somewhat clearer picture of the changes that have taken place in the domain of machine translation over time. Some terms have already been mentioned and described, but not all. These will be explained in more detail in the next section *Paradigms of machine translation*.

3.2.2 Paradigms of machine translation

This section describes scientific theories and models that have been applied in machine translation over time. These theories and models formed the conceptual framework that were adopted at the time. Some of these models and / or theories have already been previously mentioned in the section "History of machine translation". The purpose of this section is to provide a comprehensive picture of the views of experts who have had a major impact on the approaches that have been and are still being used. It also helps to better understand translation machines and the processes involved in them.

3.2.2.1 Rule-based machine translation

In machine translation, systems can be split in *knowledge* or *data driven* systems. Rule-based translation machines were the first type of translation engine to be used and these are the only *knowledge driven* systems. Rule-based translation machines are systems in which dictionaries with common words are combined with linguistic rules. The translation engine must be fed with the user's dictionaries to improve the translations. As a result, the result will not immediately meet the end user's expectations. Nevertheless, rule-based translation systems can usually produce coherent and logical translations if the right specialized dictionaries are used. In rule-based machine translation, three different approaches can be considered: *dictionary-based*, *transfer* and *interlingual* approaches.

In the very first translation systems, up to around 1966, the source language was converted into the other language, the target language, with as few steps as possible. This "direct approach" was based on a word-for-word translation, in which the environment of the word in the source language was only considered a choice between various words needed to be made in the target language. For example, when translating "fly" into French, it is important to know whether it is a verb or a noun ("I fly" = "je vole", "a fly" = "une mouche"). A complete morphological and syntactic analysis is therefore not used as an intermediate stage for the translation in a direct translation system; parts of speech are only determined when that information is "really needed". The resources that are used for this approach are generally limited to a bilingual dictionary, providing target language word equivalences (Hutchins, 1978). Therefore, this approach is also defined as a *dictionary-based approach*. An example of a system with a direct approach is TAUM.

The opposite of this direct approach is the *indirect approach*. Two main types of approaches are considered indirect: the *transfer-based approach* and the *interlinguistic approach*. The *transfer-based approach* is based on a deep analysis of the source text that operates over three stages: analysis, transfer, and synthesis (also called generation). "Analysis", a component with source language rules, turns the source language text into a representation that is easier to translate. "The representations are language specific: the source language intermediate representation is specific to a particular language, as is the target language representation" (Hutchins, 1992). "Transfer" is a component that translates the source text into a rudimentary target language, whose form is then manipulated by the "Synthesis" component. Each phase of the process uses specific dictionaries.

In an interlingual system, analysis and synthesis are so extensive that all transfer rules are superfluous. The *interlingua*, according to Hutchins (1992), "is an abstract representation of the language, it includes all information necessary to the generation of the target text". This means that the "transfer" component is not necessary. Proponents of an interlingual system often justify their preference by saying that an interlingual translation system is efficient: in an interlingual system for four languages, you need four analysis and four synthesis components; in a transfer system an additional twelve (4 × 3) transfer components are needed. The number of transfer components increases dramatically with each new language added to the system.

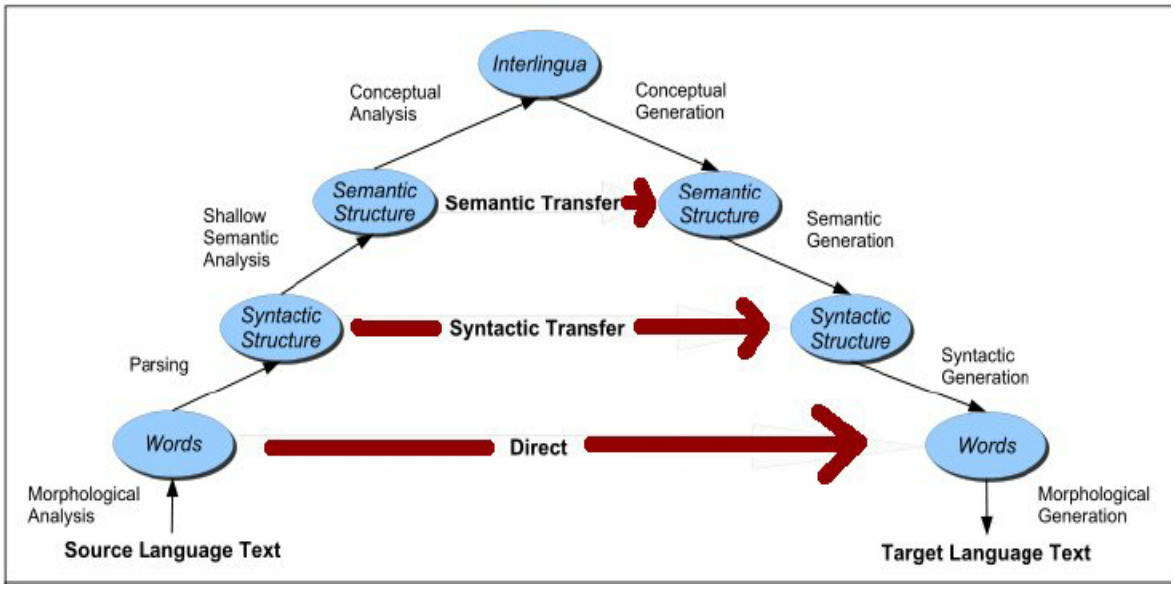

Figure 9: The Vacquois triangle (Gompel, 2009)

In the figure above all three systems are summarized graphically. It makes apparent that the systems differ in the degree to which analysis is performed before translation takes place. The left side represents the component "analysis" and the right side the "synthesis" component. In the first layer the "direct approach" is shown, an approach based on a word to word-to-word translation and in which analysis is limited to the lexical level. In the middle, the *transfer-approach* is represented. In this approach there is an analysis at syntactic and semantic level before this information is transferred to the "generation" of the target text. The top represents the *interlingua-approach* that analyses the source and target text at all levels, to generate the translated text.

3.2.2.2 Example-based machine translation

Another approach that can be distinguished is the *example-based approach*. This is one of the first approaches that can be considered a *data driven* system. Example-based machine translation is based on the idea of analogy, an approach that was proposed by Japanese computer scientist Makoto Nagao in 1984 (Nagao, 1984). With his new approach, Nagao tried to solve the weaknesses of *rule-based* translation machines: according to this author, when translating between languages with completely different structures, such as English and Japanese, there is no use of deep linguistic analysis. Nagao's and other *example-based* systems use segments of the source language, extracted from a large corpus, to build texts in the target language with the same meaning (Hutchins, 2005). Thus, "the main idea of this approach is to find matches (correspondences) among words, with the aim of achieving the best option between the source language and the target language, by using texts that were already translated by other translators" (Testa, 2018). To get a better picture of the processes in *example-based* machine translation, a simplified model is shown below:

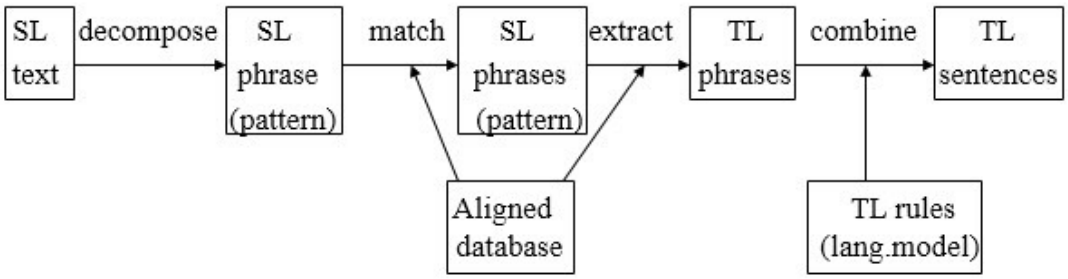

Figure 10: An example-based translation model (Irfan, 2017)

The process in example-based translation machines can be divided into approximately 3 steps: matching, alignment, and recombination. First, when give a source language sentence, the translation system compares it with source language sentences that are stored in the corpus, trying to find a match, chooses which examples are the most useful for translation and stores them. After that, in the "alignment phase", it identifies which parts of the corresponding (matching) translation are to be reused. "Recombination" is the final phase. In this phase, the machine makes sure that all segments selected during the alignment are put together in a legitimate way and reorders them into translation units.

3.2.2.3 Statistical machine translation

Example-based machine translation, was replaced relatively quickly by *statistical machine translation*. Example-based machine translation can be seen as a kind of building block for statistical machine translation because they are similar in several ways. For this reason, both systems are sometimes named *corpus-based* systems. For example, both example-based translation and statistical translation have large bilingual corpora as their fundamental data source (Somers, 1999). Additionally, neither involves deep linguistic analysis, because their developers, such as Nagao, do not see the use of it when translating from and to languages that are completely different.

Yet, despite having similarities, there are also significant differences. According to Hutchins (2005), *statistical translation* machines extract individual words while *example-based translation* machines extract segments (instead of individual words), as is explained before. Also, statistical translation machines use statistical data (such as parameters) derived from corpora data, and thus preprocessing the data is essential in this type of translation engines. In *example-based translation* machines, preprocessing the data is optional and corpora are used as a primary data source.

The original idea of *statistical machine translation* was introduced by Warren Weaver, although Weaver never developed a statistical translation machine. His ideas were later re-introduced by IBM, resulting in the first statistical translation machine CANDIDE (mentioned in the "history" section). It is also known that Google made use of this approach. In 2005, it used a 200 billion-word corpus of United Nations documents to train their system, causing a large improvement in translation accuracy (Google, 2005). Before the introduction of neural translation machines, *statistical machine translation* was the most widely studied and implemented translation method.

3.2.2.4 Hybrid translation machines

After a period of extensive use of *example-based* and *statistical approaches*, some researchers developed hybrid translation systems. The period of *hybrid translation* machines is not indicated on the previously shown IBM timeline, but it is nonetheless interesting to mention here.

Hybrid translation systems arose from the idea that all translation problems could not be solved by a single method and guarantee good translation quality. Until 1990 specific systems used a single method or approach (for ex: *rule-based*, *example-based* etc.). Hybrid translation machines are thus a combination of methods that were already being used. However,

different combinations are possible. For example, the so-called *multi-engine* system of the Carnegie-Mellon University Group (2005) was a combination of *rule-based* and *example-based* systems. Another example of a hybrid system was Microsoft's (Dolan, Pinkham, & Richardson, 2002), which represents the most common combinations in hybrid systems. In this system, statistical methods from *statistical* and *example-based translation* machines are combined with linguistic-based methods from *rule-based translation* machines (Hutchins, 2016).

Hybrid systems, through a combination of systems that already existed at the time, aim at the extracting the best features of each approach, to provide the best translation quality, allowing exploration and improvement of both systems.

3.2.2.5 Neural translation machines

The most recent type of translation machine is the *neural translation* machine. *Neural Machine Translation* (NMT) is a form of machine translation that uses a large artificial neural network to make predictions about the probability of a sequence of words. The first scientific paper in which the use of neural networks in machine translation was proposed appeared in 2013 (Kalchbrenner & Blunsom, 2013). This paper describes a model that will "encode a given source text into a continuous vector using *Convolutional Neural Network* (CNN), and then use *Recurrent Neural Network* (CNN) as the decoder to transform the state vector in the target language". This work is considered as "the birth of neural machine translation". After that, *neural translation* machines became rapidly popular and successful. In WMT'15, an annual translation machine competition, a neural network-based translation machine appeared for the first time. In 2016, there were 90% of neural machine translation systems among its winners (Bojar, et al., 2016).

As mentioned in the first paragraph, *neural network-based translation* engines have an encoder-decoder architecture. The encoder's neural network reads and encodes a source sentence into a vector, a sequence of numbers representing the meaning of the sentence. A decoder then performs a translation of this encoded vector. Initially, as stated earlier, mainly the so-called *RNN* and *CNN* networks (which are types of neural networks) were used, but both had weaknesses. *RNN* would be suitable for dealing with smaller segments, while *CNN* would be more suitable for longer segments. To solve this problem, a so called *attention mechanism* was introduced (Bahdanau, Cho, & Bengio, 2016). The *transformer architecture*, an *attention-based* model, is the most common encoder-decoder architecture nowadays (Barrault, et al., 2019).

Yet there are companies that still use, for example, *recurrent neural networks*. An example of such a company is Google, whose translation machine Google Translate has its own architecture. This architecture is represented in a common sequence to sequence model with an attention mechanism and a *long short-term memory architecture* (a type of recurrent neural network). In the figure below, Google's architecture is schematically presented.

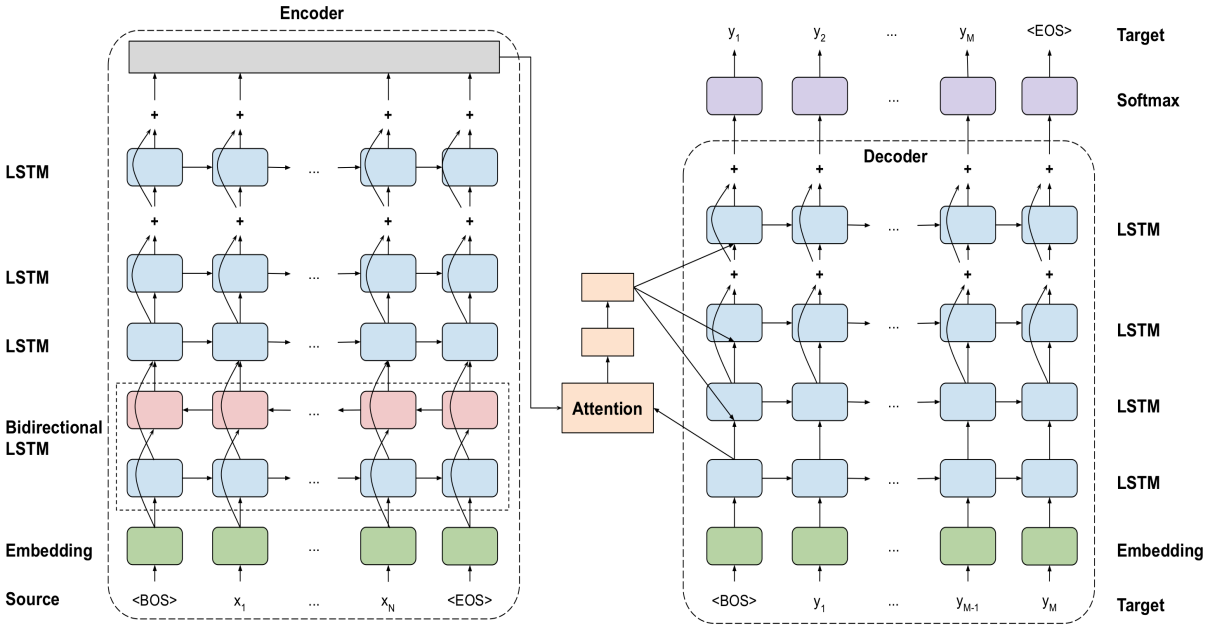

Figure 11: Googles neural network translation machine architecture (Shrestha, 2019)

3.3 Quality Assurance

As mentioned before, ensuring quality of service is considered important for Unbabel. *Quality assurance* is a well-known term in the business world, designating methods and strategies used by companies to guarantee the quality of their services. This includes translation agencies or, in Unbabel's case, translation platforms. To achieve this, quality management systems or models can be used. As briefly mentioned in the "The Company" section, examples of such methods and strategies used at Unbabel involve tools such as the *Smartcheck* or the *Dependency Parser*. For the remainder of the thesis, it is relevant to go into a little more detail and provide some background information concerning these quality assurance models, especially the models that are adequate for assessing translation quality.

3.3.1 ISO

ISO is an international organization consisting of a partnership of national standardization organizations in 163 countries that sets standards and values. These standards are written down in documents defining required specifications, guidelines, or characteristics (ISO, 2017). These can be used by companies to ensure that materials, products, processes, and services are fit for purpose and to ensure worldwide quality, safety, and reliability.

ISO standards that are important in the translation world are the ISO 9001 and 17100. These are used for quality management and focus on, among other things, revenue growth, proven quality, higher customer satisfaction, efficiency, and cost savings. Many translation agencies are ISO 9001 and 17100 certified, although this is not the case of Unbabel.

3.3.2 LISA

The first example of a *Quality Assurance* model in translation is the LISA model. It was developed and disseminated in 1995 by the *Localization Industry Standards Association* for language localization projects that has been applied to product documentation, user interfaces and even computer-based training (e-learning) (Parra, 2005). It includes a predefined list of error levels based on severity and relevance, an overview of error categories, a catalog of the reviewer's tasks and a template to indicate whether the translation was successful or failed. In the image below, a visual representation of the LISA model is given.

Figure 12: Example of LISA QA Model Interface (Localization Industry Standards Association, 2006)

There are three severity levels of error: *critical*, *major,* and *minor*. The more serious the error, the higher the number of points allocated to reflect the severity of the mistake. Minor errors are the least serious, major errors are in the middle of the scale, critical errors being the worst. These are respectively worth 1, 5 and 10 points. Since critical errors represent the most serious type of error, even if only one of these is found, the translation/localization fails immediately. Critical errors are worth 1 point more than the maximum number of error points allowed (Parra, 2005).

The maximum error points allowed within each category are calculated automatically using the number of words translated. The "total" column counts the number of points scored in each section. This column provides information about which error categories are most problematic, making problem areas clearly identifiable.

During the translation revision, the number of errors are entered in the corresponding fields according to the error category and severity of the error. As the errors are entered in the

form, a PASS or a FAIL appears automatically in the "Result" area, represented in green or red, respectively.

Since 2011, LISA is no longer active. It is said that LISA did not keep up with the times and lacked the flexibility that is required in a world with diversified types of content. (Görög, 2017). Nonetheless, its standardization methods continue to be widely used in translation quality evaluation.

3.3.3 TAUS

Another example of a quality assurance model is the *Dynamic Quality Framework* (DQF), developed by the Translation Automation User Society (TAUS) in collaboration with Sharon O'Brien (2012). The framework consists of a set of tools that seeks to evaluate both human and machine translation. It covers several features and competences, including accuracy, fluency, evaluation based on error-typology, productivity measurements, content profiling, and a knowledge base of best use cases and practices. By means of an open API, a software interface that enables communication between two applications, users such as translation purchasers, project managers and freelance translators can monitor the quality, productivity, and efficiency of the translations. The API thus connects the translation tools with the DQF.

As mentioned in the first paragraph, the framework offers an evaluation based on error typology. A vast majority of buyers and providers of translation services manage their quality program through this (Lommel, et al., 2015). In the 80's the LISA model formed the basis of most error typologies. TAUS, however, tried to develop a more up-to-date typology through the DQF. As the LISA model, it involves a list of error categories. The content of the translation is reviewed by a professional linguist who detects and points out errors and determines whether it can proceed or should be rejected.

Important differences in comparison with the LISA model, for example, are the definition of the error categories and the severity levels. The tables below show how TAUS defines the main error categories and severity levels.

Language	Although it can refer to ambiguous sentences, an error in this category generally means a grammatical, syntactic or punctuation error.
Terminology	A glossary or other standard terminology source has not been adhered to.
Accuracy	Incorrect meaning has been transferred or there has been an unacceptable omission or addition in the translated text.
Style	Quite subjective, it refers to a contravention of the style guide.

Figure 13: Definition of main error categories (Lommel, et al., 2015)

Severity 1	Critical errors may carry health, safety, legal or financial implications, violate geopolitical usage guidelines, damage the company's reputation, cause the application to crash or negatively modify/misrepresent the functionality of a product or service, or which could be seen as offensive.
Severity 2	Major errors that may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or because errors appear in a visible or important part of the content.
Severity 3	Minor errors that don't lead to loss of meaning and wouldn't confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Severity 4	Neutral, used to log additional information, problems or changes to be made that don't count as errors, e.g. they reflect a reviewer's choice or preferred style, they are repeated errors or instruction/glossary changes not yet implemented, a change to be made that the translator is not aware of.
Kudos	Used to praise for exceptional achievement.

Figure 14: Definition of severity levels (Lommel, et al., 2015)

In 2015 the DQF of TAUS harmonized with the MQM model (Lommel, et al., 2015), a model that is used at Unbabel. This will be discussed in more detail in the next section.

3.3.4 MQM

As indicated earlier, the TAUS model was merged in 2015 with the MQM model, which is a metric developed by the QT Launchpad project, European Commission-funded research that intended to overcome quality barriers in machine and human translation (QT21, 2012). Arle Lommel, who worked for the German Research Centre for Artificial Intelligence and was part of this project, states the following:

> *We are especially glad to have worked with TAUS on this harmonization effort because it reduces industry confusion about which framework to use and it simplifies the implementation process for everyone. Both MQM and DQF had to make significant changes, but the resulting shared framework is clearer and more useful for everyone. Moving forward we can expect to see more industry uptake of quality assessment best practices based on this shared resource* (TAUS, 2015).

Thus, the current MQM model is an improved version of recent MQM versions and the DQF from TAUS. It provides a framework for describing quality metrics used to assess the quality and identify specific issues and errors in translated texts. The core of the model is graphically represented in the image below.

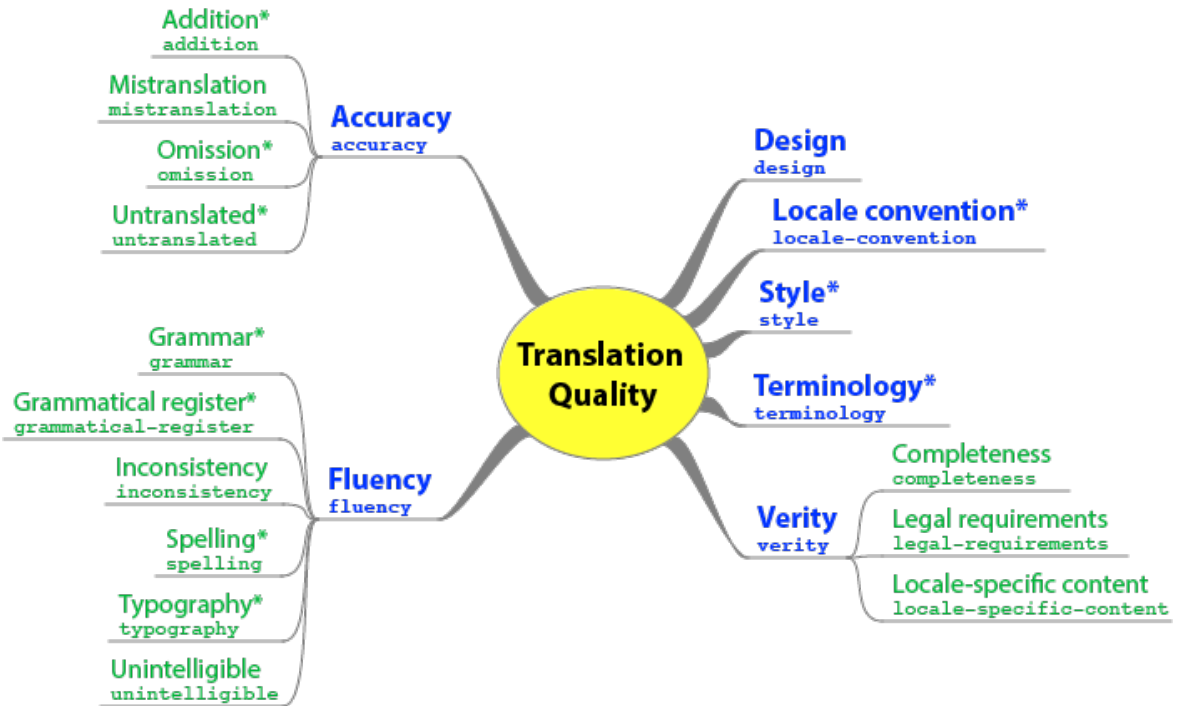

Figure 15: Graphical representation of the MQM model (Liu, 2018).

The framework consists of a vocabulary for categorizing quality issues (which is the core represented in the image above). The vocabulary for categorizing quality issues serves well for companies that make use of an annotation process (such as the one set in place at Unbabel, which has been described earlier). This way, an annotator can accurately define what type of errors are made. As is visible in the graphical representation, there are several error categories, such as design, accuracy and fluency, and subcategories, such as grammar, omission and typography. The QT Launchpad project further describes these issues in an online document (QT21, 2015).

In addition, it contains a scoring mechanism to provide quality scores based on counts of errors or error annotations. Like the LISA model, the MQM model defines errors as "minor", "major" or "critical. In the original MQM model, the following values are assigned: 1 for a "minor" mistake, 10 for a "major" and 100 for a "critical". The scoring system determines quality on the basis of the following formula: "TQ = 100 - TP + SP", where TQ stands for quality score (the general quality assessment), TP for penalties for the target content (sum of the assigned values to the target text) and SP for penalties for the source content (sum of the values assigned to the target text) (Lommel, et al., 2015). The higher the score (with 100 as the highest possible), the better the quality of the text.

Also, the framework has a set of guidelines for selecting the issues. These guidelines contain indications on which error category to be used to define an error. Like the vocabulary, these are further described in the online document of the QT Launchpad.

As mentioned in "The Company" section, at Unbabel quality is very important, and producing high-quality translations is one of Unbabel's objectives.

> *Our objective is to have all the LP's (language pairs) assessed (90% of the content of our pipeline) at a 95 MQM, perceived in the industry as the professional level* (Oliveira, 2019, p. 6).

Unbabel uses the MQM model to improve its translation services. To do this the vocabulary, scoring mechanism and the guidelines are part of the translation, post-editing and annotation processes. Being so, many translations in several language pairs already have been scored and annotated. However, it should be mentioned that Unbabel adapted the error typology used, which means that it is slightly different from the original MQM typology. This is shown in the company's guidelines, like the one provided in the appendices. Apart from that, Unbabel uses the same features as the original MQM model. In addition to the fact that Unbabel can determine the quality of the target text by identifying errors through this model, this model is also a useful

tool for making improvements, as the company can, for example, check which are the most common errors and outline possible solutions. This is an important advantage, as it makes work such as the one presented in this thesis more easily feasible. This will be further explained in the “Methodology” section.

4. Methodology

In this chapter we provide general information regarding the research developed under the scope of this work. For example, it describes what kind of research it is, how it was set up, its goals and who or what contributed to the research.

4.1 Research Questions

Firstly, we established research questions that provide structure and organization to this thesis and indicate what is specifically meant to be researched. Thus, one of the main goals of the research developed is to provide answers to these questions.

4.1.1 Main question

- To what extent error patterns can be found in the performance of the automatic speech recognition (ASR) and machine translation (MT) systems used at Unbabel and how do they influence each other?

4.1.2 Subquestions

- How do the ASR and the MT systems perform?
- How could the errors found be defined?
- What differences can we find per video type, if any?
- How can Unbabel eventually deal with the error patterns identified and improve its performance?

4.2 Research goal

The research questions introduced above show what we intend to achieve by conducting this research. First we should mention that this work aims at helping Unbabel to improve the quality of its services. More specifically, this research serves to try to gain new insights into error patterns and the performance of the ASR and MT systems used at the company. Within the scope of this line of work, it is also important to examine to what extent the ASR has an influence on the performance of the MT system. By examining these error patterns and performance, and with the answers found through this study, Unbabel will be able address the issues identified and improve the quality of its services.

4.3 Research Design

In this chapter we describe the type of research conducted, how the data was collected and how the project was developed.

4.3.1 Type of research

Our research questions are a combination of defining questions (how can the errors be defined?) and exploratory questions (how do the systems perform, what patterns can we find?). Therefore, we can consider that the work presented here consists of *exploratory research* and *observations*.

Exploratory research is research in which the examiner systematically collects and analyzes data trying to discover new relationships or acquiring new facts (Hulp bij Onderzoek, 2017). It is conducted to have a better understanding of the existing problem but often does not provide conclusive results. Thus, this work serves as preliminary research or groundwork for future studies. Any factors that might be relevant regarding the subject, possible relationships and underlying motivations are examined, and tentative conclusions are drawn, so that other research subsequently can build on this research and provide clearer statements. The fact that exploratory research helps to lay the foundation for future works and that it helps to get a better insight regarding the existing problem are considered one of the advantages of this type of research (Formplus, 2007). Disadvantages of this type of research are that it often does not provide definite conclusions, as mentioned.

Under the scope of our research observational data collection methods were used. These simply refer to methods in which a certain behavior (in this case the behavior of the ASR and of the MT systems) is observed (Dingemanse, 2018). They can be qualitative or quantitative, and for this research we opted for an *unstructured observation* method. Unstructured observation is a method that does not use a pre-established observation schedule, but instead collects as much information as possible in which behavior is observed and described in detail. This means that you can form a broad picture of the situation to be investigated and the behavior to be investigated, because you do not focus on a certain element, but on everything. The advantage of not having a pre-established schedule is that you can conduct a broader observation to identify key aspects of the problem and, for example, formulate hypotheses (Dingemanse, 2018). The disadvantage, however, since there is no pre-established schedule, is that there is a great risk that you (unintentionally) focus unnecessarily on a certain aspect and on misinterpretations or non-scientific interpretations.

Despite the aforementioned disadvantages, it was decided to opt for these types of research and data collection methods, because, after consultation with Unbabel, they align better with the problem description and objectives. The research design and how the analysis was conducted is described further below.

4.3.2 Research approach

To get a better picture of the performance and the errors produced by the ASR and the MT systems, first, data was generated from videos whose speech was automatically recognized and then automatically translated. After consultation, it was determined to generate data from 10 videos of 1-3 minutes to be able to gather a reasonable amount of information to detect eventual error patterns and examine the performance of the systems.

Amount of videos	**10**
Duration videos	**1-3 minutes per video**
Type of videos	**User content, Professional content**
Language pair	**English (US/UK) – Dutch**

Figure 16: Description characteristics videos

The type of videos that Unbabel usually translates include *user content*[2] and *professional content*[3]. To be able to find if there is a link between the type of video and the errors of the ASR and of the MT systems, we decided to gather data from both types of videos, 4 videos containing user content, and 6 videos containing professional content.

It was also decided to generate data from videos that were translated from English into Dutch. This is because of some reasons. First, because Unbabel translates many from English into another language, and second, Dutch is a language that has not been examined in Unbabel so far. UnBabel chose a native-Dutch examiner because it considers that native speakers are generally more suitable to find patterns in errors made in their native language rather than someone who investigates a language that is not his or her native language. Regarding the source language, Unbabel opted for videos in both American-English and British-English,

[2] Refers to any type of content information created by users belonging to an online platform (Kang, 2019).
[3] Refers to any type of content information that is business-like or provides necessary information to improve the jobs of professionals (van Bregt, 2012).

because in that way it is also possible to find out if there is a connection between the type of English and the errors made by both systems.

4.3.3 Data collection procedure

As mentioned above we analyzed the automatic transcriptions and the automatic translation of 10 videos. To be able to get an idea of the errors made by the ASR and MT systems used at Unbabel, the text spoken in the 10 videos was transcribed and then provided in Word files. These word files contained the automatically transcribed text, the automatically translated text, but also the transcriptions already made by humans. In this way, the detection of errors should be easier, thus saving time. When working with videos, it is also often useful to have the footage at hand, and this was the setup in which the research presented in this thesis was developed: besides the information in the Word files mentioned above, a set of links to view the videos themselves was made available.

4.3.4 Data analysis method

The analysis of the data was done using the annotation process described in the "The Company" and "Literature Review" chapters. We performed our own annotations and by using the error typology described in the "Annotation Guidelines" the errors were identified and labeled.

All errors identified in the annotation process are listed in an Excel file, which can be found in the appendices. This file contains all data and relevant information regarding the transcriptions and translation of the videos. For example, there are separate columns for the automatic transcription performed by the ASR, the transcription performed by a human transcriber and the automatic translation. The column "Remarks" explains what the error is and why. For clarification, we suggested our own translation in the column "Opted Translation". On the right side of the file, you will find the typology of both the ASR and MT errors for each error. In addition, to detect any patterns, it is indicated whether each translation error was caused by the ASR or not, and if so, by what type of error. At the far right it is indicated whether the errors have been considered as *minor*, *major* or *critical*.

We also looked at the *lexical density* of speech of the videos. What this means is explained more explicitly in the "Analysis and Results" chapter. The lexical density per line is shown the "Lexical Density" part in the Excel file.

All data shown in the Excel file is extensively described in the "Analysis and Results" chapter of this work. To meet the aims of this work, we decided to elaborate on each error type and error combination. This was done based on tables with examples, descriptions and possible findings, such as the relationship between the ASR and the MT systems. Percentual data was also provided to get an idea, for example, of which errors and error combinations occurred the most. Finally, the chapters "Findings and Discussions" and "Conclusions" discuss the findings as a whole and the conclusions that could be drawn from this research.

4.3.5 Video descriptions

It can be important to provide a description for each video as it is possible that the performance of the ASR and the MT systems will differ due to specific characteristics of the videos. First a general description is provided, followed by a more specific description of each video.

The first five videos are videos from the American media company Great Big Story, a business that was launched by CNN in 2015. Great Big Story makes microdocumentaries and short films that are often viewed on social platforms such as Facebook and YouTube. In these videos there are two speakers: the interviewee and the narrator. The videos have a formal atmosphere. We should also mention that in some of these videos, the speakers do not seem to have English as their native language, a fact which can eventually lead to different performances of the ASR and the MT systems. It is likely that the ASR will be the most affected, since errors in transcription generally impact the machine translation.

The sixth, seventh and eighth videos are videos in which general public individuals, mainly of young age, give their opinion about certain products. This kind of videos can be called *user content*. In these videos, there is only one speaker and the register that is used is very informal. In one of the videos, the speaker is British. This could also possibly have an impact on the performance of the ASR and MT, as has been proven that British accents are harder to recognize for ASR systems.

The ninth video can also be considered a *user content* video, but differs from the other three as there are several speakers. The register used in this video is slightly more formal than the one used in the previous three.

The last video is a video of media company CNN, which can be considered as *professional content*, but differs a bit from the other videos with professional content.

This video is not a microdocumentary, but a small news broadcast. This video has several speakers.

4.3.5.1 Video 1

In the first video, an entomologist named Dr. R. Isaí, explains how he tracks down new species of insects in the Patagonian Ice Field. Due to the drastically changing climate, it is hard to get any information about the insects that inhabit the area. Despite of that, Dr. Isaí was able to discover species of insects with a different coloration on the bottom, and to be able to do that, he needs to develop some unexpected skills, like carrying his tools to new locations with a low carbon footprint. Despite of the dangers involved in the entomologists work, he is not planning on slowing it down because the home of the insects is disappearing rapidly, due to climate change and human impact.

4.3.5.2 Video 2

The second video is about a country musician from Kenya called Elvis Otieno. Otieno explains that country music was his first love and that he inherited this love from his parents. He also clarifies that he identifies himself with country music because of the message in it, such as family, love, and heartbreak and that he draws inspiration from many different artists, such as Garth Brooks, Charlie Pride, Alan Jackson, and Don Williams. The struggle that they had is universal. What the musician wants his audience to feel are the emotions of a song.

4.3.5.3 Video 3

The third video shows, Gregory Loan, a senior simulation engineer, working for the Boston Children's Hospital. Loan makes artificial patients to allow real doctors to practice procedures to improve health outcomes for kids. Something that is considered important for the engineer's work are the experiences with special effects, which he gained when he made, for example, dinosaurs for the Jurassic Park theme parks or magical creatures for Harry Potter. Next to special effects, Loan has another passion, robotics, which he uses with special effects, engineering and medicine to produce something that helps people. Mainly he makes simulation models, such as an arm that you can inject, which are used for a single goal: save children's lives.

4.3.5.4 Video 4

The fourth video explains the story of "video game player of the century" Billy Mitchell. In 1999, he got crowned like that by Masaya Nakamura, also called the godfather of video

games, after reaching a perfect score of 3,333,360 in Pac-Man and leaving the game with computer garble and without half of the memory. After starting in the competitive world of pinball, he moved to Donkey Kong and later Pac-Man, which was, according to Mitchell, the most competitive game of that time. After four to five hours of pure focus, perfect timing and without dying once, he achieved something that, in his opinion, could not get any higher: ending a game that was not even designed to end.

4.3.5.5 Video 5

The last video of the *Great Big Story* shows the passion of Pam Utharntharm, a chef from Thailand that spends almost all of her time cooking. She owns a restaurant called "The Table" in her own home, where one big table per night can be booked by her customers. What started off with family and friends, now has a waiting list of three months. Utharntharm used to work for a three-star Michelin restaurant in New York but decided to go back to Bangkok to open her own restaurant. To meet the expectations of her customers, she tries to obtain local ingredients of distinctive flavor. According to the chef, she is not successful yet and for her and other chefs it is important not to think that you are successful, because that will stop you from learning new things. Being a chef is not a job, but a passion for her, and if the customer returns home with a smile and full stomach, she is happy.

4.3.5.6 Video 6

The sixth video is a product review of the Whirlpool Duet WFW94HEXW washing machine, considered as one of the best washing machines in the industry. It is a highly rated washing machine and even rated number one in a consumer magazine (name unknown). The reviewer further explains how to use the machine. One of the reasons why the reviewer likes the machine is because the machine is really easy to use.

4.3.5.7 Video 7

The seventh video shows a man that describes and gives his opinion about McDonald's *Vegetable Deluxe burger*. It is a sesame seed bun, whose main ingredient is chickpeas, with lettuce, coriander and *sandwich sauce*, a similar sauce to mayonnaise, but slightly different. According to the reviewer, the burger is a little bit dry and lacks flavor, which possibly can be solved by adding more sauce, salt, and pepper. The predominant flavor is the one of the vegetables and the coriander, which might make the burger not so tasty for the regular public but meet the expectations of vegetarians. The overall rating given by the reviewer: 3 out 5 stars.

4.3.5.8 Video 8

The eighth video is published by a make-up reviewer called Rachel and shines a light on a 24-hour lipstick from Aldi, that she bought because of another review of a friend. According to her, the color is pretty but after seven hours it gets faded, which is not that bad. It does not look weird, but the fact that it supposedly lasts for 24 hours is not true.

4.3.5.9 Video 9

The ninth video is about a scientific experiment with wood blocks. The experiment aims at testing what would happen if a bullet were shot exactly in the center of the block and what would happen if it were shot off center of the block. The creator of the experiment previously asked the opinions of people in another video and shows in this video if they were wrong or not. There were three hypotheses: the block that gets shot in the center ends up higher than the block that gets shot off center, it ends up lower than the other block or they both end up at the same height. After the experiment, both blocks end up at the same height, which surprises many people. According to them, it would make more sense if the block that is shot off center would end up lower due to its extra rotational energy. In the end of the video, the creator of the experiment asks its viewers to think about a possible explanation for this and states that he will reveal the answer in another video.

4.3.5.10 Video 10

The last video is a video from CNN, in which a system in police cars for tracking down license plates of other cars while driving is discussed. From the perspective of a police officer, the system is a piece of technology that helps to keep people safe, but in the point of view of a local activist called Mike Katz-Lacabe, this is a violation of privacy. The video shows that the system is not limited to taking pictures of the license plates, but also covers things around it. In one of the pictures, it is possible to see Mike Katz playing in front of his garage with his kids. In an era of digital rights and privacy some people say there needs to be more transparency and limits to which information can be gathered.

5. Analysis and Results

Initially, it was decided to perform a "pre-analysis", considering the characteristics and the content of the videos. To gain a good insight of the content and the complexity of the videos that were used to assess the performance of the tools, we decided to examine the *Lexical Density*, the *Readability*, and the average length of the sentences, as was already briefly mentioned in the "Methodology" section. These were examined, because they could be a possible reason for different results in the outputs of the ASR and of the MT systems. Below we describe what we mean by these features.

5.1 Lexical Density

Victoria Johansson (2008), a student at Lund University in Sweden, describes in detail what lexical density means and how the concept has evolved over time.

The concept of *lexical density* was originally proposed by Jean Ure (1971) and describes the proportion of content words, which are words that provide most information in a sentence (nouns, adjectives, verbs, and adverbs), in the total number of words in an utterance. According to Johansson, humans receive a notion of information packaging; a text with a high proportion of content words contains more information than a text with a high proportion of function words, which are words that provide less information in a sentence, like prepositions, interjections, pronouns, conjunctions, and count words.

Ure distinguishes between words with lexical properties and without. According to the author, words without lexical properties can be described as "purely in terms of grammar" (Ure, 1971). These words have a more grammatical-syntactic function than the lexical items. In that case, lexical density can be calculated by dividing the lexical items by the total number of words, leading to the formula below, in which L_d represents the analyzed text's lexical density:

L_d = the number of lexical items/the total number of words * 100

The result obtained by applying the formula is a percentage, which indicates the extent to which a text consists of lexical words and thus is lexically dense. Ure concluded that most spoken texts have a lexical density under 40% while most written texts have a lexical density of more than 40%.

In a later article, Ure defines *lexical density* as the proportion of lexical words/items to the words with grammatical values (instead of the total number of words). In this article she also states that the matter of lexicality is important when discussing the concept of *lexical*

density. The word classes that have lexical properties are nouns, verbs, and adjectives. These items are also called *content words* or *open class words* while more grammatical parts (pronouns, prepositions) are called *closed class* words.

Later, the concept of lexical density is further developed and refined by the linguist Michael Halliday (1985). Halliday states that it is important to make a good distinction between lexical items and grammatical items. A lexical item, in his opinion could have more than one word. While Ure counts, for example, "turn up" as one lexical item and one grammatical item, Halliday counts it as one lexical item. According to Halliday, a lexical item is defined as an item that functions in lexical sets, not grammatical systems. The lexical item is part of an open set that can be contrasted with several items. In contrast, a grammatical item, is part of a *closed system*. According to Halliday, what is characteristic for the grammatical system is that the word classes that belong to it have a fixed set of items. This makes it impossible to make new word class members.

To reinforce his point of view, Halliday uses child language as an example, which proves the existence of two classes, one with lexical and one with grammatical items. In the early stages of language development of children, children often create sentences that lack lexical items. He also further emphasizes that lexis and grammar are strongly connected. For instance, he claims that English prepositions and certain types of adverbs are on the border between lexical and grammatical items. Examples of this are modal adverbs such as "always" and "perhaps". For example, when comparing spoken language with written language, it does not matter that much where you draw the line between lexical and grammatical adverbs, but what matters is the consistency in drawing it.

Thus, Halliday's definition of lexical density corresponds to the number of lexical items as a proportion of the number of running words (Halliday, 1985). The most important difference between Ure and Halliday is that Halliday counts some adverbs as lexical items. Next to that, these lexical items can be formed by more than one word, which clearly has an impact on the count.

5.1.1 Lexical vs. grammatical words

As mentioned before, Ure (1971) defines lexical density as the proportion of lexical words/items to grammatical function words. It might be important to define what function words are so that the analysis can be performed as accurately as possible. To get a clear idea of

what lexical words and grammatical function words and are, this section explains in detail how both type of words are classified.

Content words are words that carry semantic content, bearing reference to the world independently of its use within a particular sentence (Winkler, 2008). Word classes that include content words are:

- *Nouns*

  A noun, can be defined as a word to name a person, place, or thing and, in linguistics, can be a member of a part of speech which can occur as the main word in the subject of a clause (Nesia & Ginting, 2014). Examples are:

  1) Persons: *Richard, Rick, Michael, student, lecturer, Asian, European, etc*.
  2) Places: *London, England, hotel, house*.
  3) Things: *telephone, book, bed*.

- *Verbs*

  Verbs are words that express action or state of being and can be classified in three groups (Hedayatnia, 1973):

  - Action Verbs, which are verbs that express actions (*give, eat, walk,* etc.) or possession (*have, own,* etc.) and are further divided into transitive and intransitive verbs.

    1) Transitive verb: a verb that always has a noun phrase that receives the action of the verb, the called direct object. For example: *Richard <u>washes</u> his car*.

    2) Intransitive verb: a verb that never has a direct or indirect object. For example: *Michael <u>sighs</u> deeply*.

  - Linking Verbs, which are verbs that connect the subject (*Jason*) of a sentence to a noun or adjective that renames or describes the subject (*business manager*). This noun is called the subject complement. For example: *Jason <u>became</u> a business manager*.
  - Auxiliary Verbs. An important note is that this type of verb is classified as a function word (Ure, 1971). Thus, auxiliary verbs will be further characterized in the function word section.

- *Adjectives*

  An adjective is defined as a word which is used to describe a noun, whose main role is to modify a noun or pronoun, giving more information about it (Dahami, 2012). The function of adjectives in English is to add clarity to the meaning of nouns. Examples are:

  1) The tulip is a *beautiful* flower.
  2) English food is the *worst*.
  3) He ate *some* porridge.

- *Adverbs*

  Adverbs are a heterogenous group of items, whose most frequent function is to specify the mode of action of the verb (Crystal, 1980, p. 16) and may modify a verb by giving circumstantial information about the time, place or manner in which an action or a process take place (Finch, 2000, p. 84). Adverbs are used to give more information and to modify verbs, clauses and other adverbs and are mainly formed by adding the suffix "-ly" to the end of an adjective (*nicely*) (Poai, 2012). Examples are:

  1) He spoke *loudly*.
  2) That is not good *enough*.
  3) We stayed in Mariana's house *all day*.

As stated above, content words are words that carry semantic content, bearing reference to the world independently of its use within a particular sentence (Winkler, 2008). Function words, on the contrary, have a more 'non-conceptual meaning':

> *As opposed to content words, function words have a more non-conceptual meaning and fulfill an essentially 'grammatical function'; in a sense they are needed by the surface structure to glue the content words together, to indicate what goes with what and how* (Corver & Riemsdijk, 2001, p. 1).

Function words fall into the minor parts of speech, including prepositions, pronouns, conjunctions, interjections, particles, auxiliary verbs, articles, question words and some adverbs (Brinton, 2000). They are also defined as closed cases and express grammatical meaning:

- *Prepositions*

  Prepositions are words that are used to express a relation to another word or element in the clause (Harmer, 1997). They usually precede a noun or a pronoun. For example:

1) Take your sister *with* you
2) Go *down* the stairs and *through* the door
3) I put the cookie jar *on* the table

- *Pronouns*

  A pronoun is defined as a word that is used to substitute a noun, often used to avoid repeating the nouns that they refer to (Bhat, 2007). For example:

  1) Look at my dog. *He* just jumped out of the window!
  2) I am all *yours*!
  3) *That* looks like the place I used to visit when I was younger.

- *Conjunctions*

  Conjunctions are a small class of words that function as a connector between words, phrases, clauses, or sentences or as a coordinator of words in the same clause (Subrahmanyam, 2012). Common conjunctions are *and* and *as well as*. For example:

  1) Go *and* take your sister with you.
  2) God made man *as well as* the animal.

- *Interjections*

  The definition of interjections is well explained by John E. Warriner: "An interjection is a word that expresses emotion and has no grammatical relation to other words in the sentence" (Warriner, 1981, p. 128). An interjection is often followed by an exclamation mark to indicate the strength of the emotion expressed (Katz, 2019). Examples are *oh*, *ah*, *wow* and *shh*.

- *Particles*

  In English, a particle is often a small word that does not have semantic meaning on its own but relies on the word it is paired with to have meaning (McArthur, 2011). They cannot inflect, which means that their form does not change to reflect grammatical person, number, case, gender, tense, mood, aspect, or voice. Particles are very similar to prepositions and are even almost the same in terms of appearance, but there is a significant difference. Prepositions are used to create a connection between their objects and another part of a sentence, and so they have a unique lexical meaning of their own.

Particles, on the other hand, are only used to form infinitives and phrasal verbs. For example:

1) Freddy went *away* on a long trip.
2) Jimmy started *out* with sixty dollars.
3) He wanted *to* go to the beach.

- *Auxiliary verbs*

  Auxiliary verbs (also called helping verbs) are verbs that are used before action or linking verbs to convey additional information regarding aspects of possibility (*can, could,* etc.) or time (*was, did, has,* etc.) (Hedayatnia, 1973). They express tense, aspect, modality, voice, or emphasis. Auxiliary verbs accompany other types of verbs. For example:

  1) I think you *should* study harder to master English.
  2) You *may* choose what you like.
  3) Nick *will* drive to Lisbon tomorrow.

- *Question words*

  Question words (also called interrogative words) are words that introduce a question that cannot be answered with *yes* or *no* (Joshi, 2013). Common question words are *who*, *what* and *why*. For example.

  1) *Why* are you doing that?
  2) *Who* stole my money?
  3) *What* do you mean?

- *Articles*

  The last type of words that are considered function words are articles. The only articles in English are *the (definite)* and *a* (indefinite) (Yule, 1999).

For the analysis of lexical density, it was decided to use Ure's method, rather than Halliday's. This is because, Ure's method clarifies better what she considers content words and what are function words. Next to that, despite of the fact that Halliday's method is more recent, both methods correlate strongly (To, Fan, & Thomas, 2013). Finally, Ure's method is a bit more

convenient to use and more easily interpretable, because it has a percentage rather than a digit, like Halliday's method does.

5.2 Readability

Julien B. Kouamé (2010) explains in detail the concept of *readability* is. The *readability* of a text depends on its content and to test it, several aspects are to be taken in consideration, such as speed of perception, perceptibility at a distance, perceptibility in peripheral vision, visibility, reflex blink technique, rate of work, eye movements and *fatigue* in reading. Readability tests are indicators that measure how easy a document is to read and understand. Kouamé (2010) also states that readability tests can increase the validity and credibility of the evaluator, which he illustrates in his paper. There are several formulas to calculate the readability of a text. Examples of such formulas can be found in George Klare (1963) and Edgar Dale & Jeanne Chall (1949). In our work, we use the formula put forth by Peter Kincaid and Rudolf Flesch (1975), also called the *Flesch Kincaid Readability Test*, because their readability test is considered as the most reliable and most tested (DuBay, 2006).

In (Hensel, 2014), is explained how this test, officially called the *Flesh Reading Ease,* works and what the results mean. By means of a score, the test indicates how difficult a text in English is. To determine this score, a formula is executed in which two variables are key: the sentence length and the average of syllables per word. The formula represents both variables and is as follows: 206.835 – 1.015 x (total amount of words/total amount of sentences) – 84.6 x (total amount of syllables/total amount of words).

Rudolf Flesch (1949) himself explains how the scores should be interpreted. The score of the Flesch Reading Ease reflects the school years of the American educational system. To give a clearer overview of the way the scores should be interpreted, a table, that can also be found in Flesch's work, has been added below.

Reading Ease Score	Interpretation	Estimated Reading Grade
0 to 30	Very difficult	College graduate
30 to 40	Difficult	13th to 16th grade
50 to 60	Fairly difficult	10th to 12th grade
60 to 70	Standard	8th and 9th grade
70 to 80	Fairly easy	7th grade
80 to 90	Easy	6th grade
90 to 100	Very easy	5th grade

Figure 17: Interpretation of the reading scores with estimated reading grade by Flesch (1949)

As shown in the table, the texts with a higher score are of a simpler nature. To clarify even more how the scores should be interpreted, Flesch makes the comparison with formal education. For example, texts with a score between 70 to 80 are not just *fairly easy*, but they are also texts that, in general, are readable for children of the 7th grade. In an educational system such as the Portuguese, for example, that also has 12 school years, it is clear how the scores can be interpreted. The only difference is that the 13th to 16th grade represent university studies, corresponding to the three years of a university degree in Portugal. According to Flesch, most writers aim for a score between 60 to 70, which is a standard difficulty, so the text is most likely to be easy to read for most of the readers, but not too simple either.

5.3 Analyzing Lexical Density

In the Excel file in the appendix, the *lexical density*, *readability*, and average sentence length of the videos are shown. These possibly could all influence each other. In relation to the lexical density there are various assumptions that can be made, and this section shows whether these assumptions are correct or not.

5.3.1 Lexical Density and Average Sentence Length

The first assumption one could make is that there is a link between lexical density and average sentence length. The longer the sentences, the more lexically rich they should be. The scatter graph below shows whether there is a relationship between the average sentence length and the lexical density in our data.

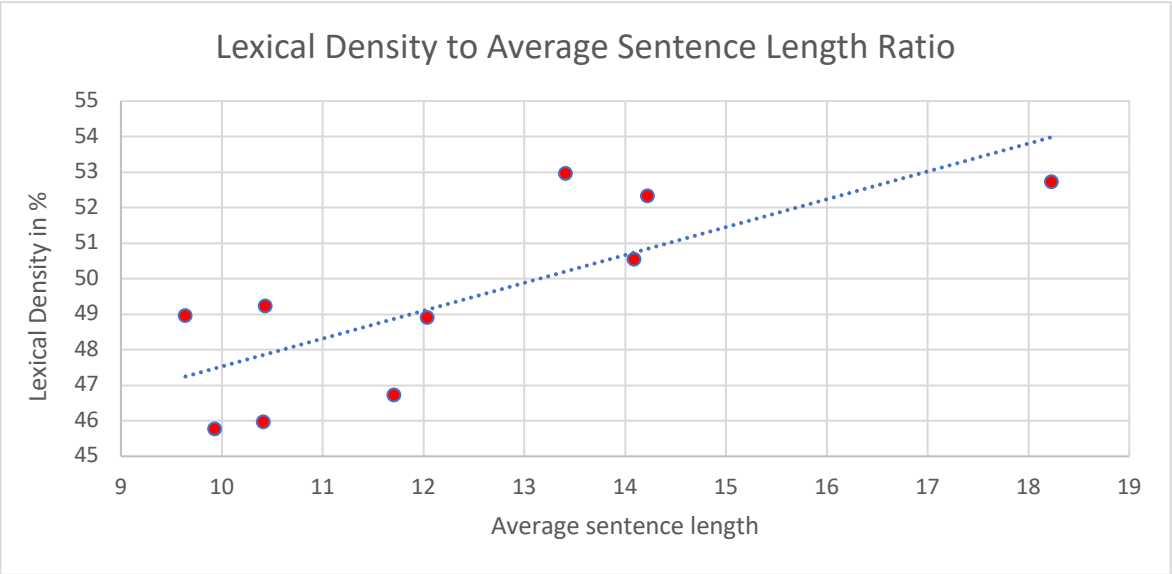

Figure 18: Lexical density in relation to the average sentence length of the videos

To be able to conclude from a scatter graph whether there is a relationship between two variables or not, there needs to be either an upward pattern (for a positive relationship) or a downward pattern (for a negative relationship) from the left to the right (Rensink, 2017). If all the dots are exactly on the regression line (the blue line in the graph), it means that there is a perfect correlation between the two variables considered. In our data, even though some dots are far away from the regression line, there is an upwards pattern from the left to the right in this graph. This indicates that there is a relationship between the average sentence length and the lexical density.

Another way of finding out a relationship between variables, is to apply a statistical test. Because both me and Unbabel decided that statistical testing is not the priority in this work, no elaborate calculations will be shown. However, we applied a simple statistical test (an independent-samples t-test) to further verify if there is a relation between the variables considered.

For such a test, according to the statistical testing steps described by Emmert-Streib (2019), it is first necessary to formulate the null hypothesis H0 (a general statement that there is no relationship between two variables) and an alternative hypothesis H1 (a statement that describes that there is a relationship between two variables in a study. In our case, the hypotheses could be defined as follows:

H0: There is no significant relationship between lexical density and average sentence length.
H1: There is a significant relationship between lexical density and average sentence length.

To find out whether there is a significant relationship (H1) or not (H0) between the two variables considered, one must find out if the p-value is lower or higher than the significance level. The significance level is often set at 0.01 or 0.05 and if the p-value is lower than these values, it means that the results are statistically significant and that H0 should be rejected (McLean, 1998). In our data, the p-value is 0.009, thus lower than the significance level, which means that the results are significant and confirms again that there is a relationship between lexical density and average sentence level.

5.3.2 Lexical Density and type of spoken English

Another assumption one could make is that the videos whose speaker is not a native speaker of English have a lower lexical density. People who are not native speakers of the language they are using tend to use a simpler language (Gürbuz, 2017). By using the data of the "Methodology" section, which describes which videos are spoken by someone who is not an English native speaker and who is, we can verify whether there is a link between the fact that the speaker's native language is English or not, and lexical density. However, as there are only three videos with non-native speakers and seven videos with native speakers, it is difficult to provide a graphical view of the differences in lexical density. Instead, the lexical densities are shown below in a table. In the table we also contrast British-English and American-English speakers.

EN (non-native)			**EN (US)**					**EN (UK)**	
Vid 1	Vid 2	Vid 5	Vid 3	Vid 4	Vid 6	Vid 9	Vid 10	Vid 7	Vid 8
52.96%	46.73%	48.92%	52.74%	45.97%	50.56%	48.97%	52.34%	49.24%	45.77%
Avg: 49.54%			**Avg: 50.12%**					**Avg: 47.51%**	
Total avg (non-native): 49.54%, Total avg (native): 49.37%, Total avg (both): 49.42%									

Figure 19: Lexical density scores per video in relation to type of English spoken

According to the data in the table, the fact that the speaker is an English native speaker or not does not seem to matter that much. The average lexical density of videos with non-native

speakers is really close to the total average (49.42%), even a bit higher, meaning that the videos with non-native speakers overall has a higher lexical density. If we compare the lexical density of the videos with non-native speakers of English with the videos with American-English speakers, the lexical density is lower, but when compared to the British-English speakers it is 2 points higher. Considering the unbalanced amount of video samples for each of the cases, we consider that there is not enough data to determine if there is a significant relation between lexical density and type of English speaker.

5.3.3 Lexical Density and type of video

Yet another assumption one could make is that there is a relation between lexical density and the type of video. For example, we can consider the hypothesis that the videos that have professional content, such as news broadcasts, have a higher lexical density. User content on the contrary, like personal reviews of a product or make-up videos, could possibly have a lower lexical density, since user content like make-up videos and product reviews are meant for a younger public, thus being more likely that speakers will use interjections such as *yeah*, *ehm* or *wow*, which are considered function words. Thus, a high number of interjections could lead to a lower lexical density. Looking at the videos described in the methodology section, there are four videos that can be considered user content videos and six professional content videos. The ones that are published by news channels, like CNN and Great Big Story (daughter company of CNN), could be considered as professional content videos, while videos published by online platform users could be considered as user content videos. The graph below compares the type of videos with their lexical density. The red dotted line shows the average lexical density of both types of videos and the columns represent each specific video. The videos are sorted from low to high lexical density, so the graph is easier to interpret.

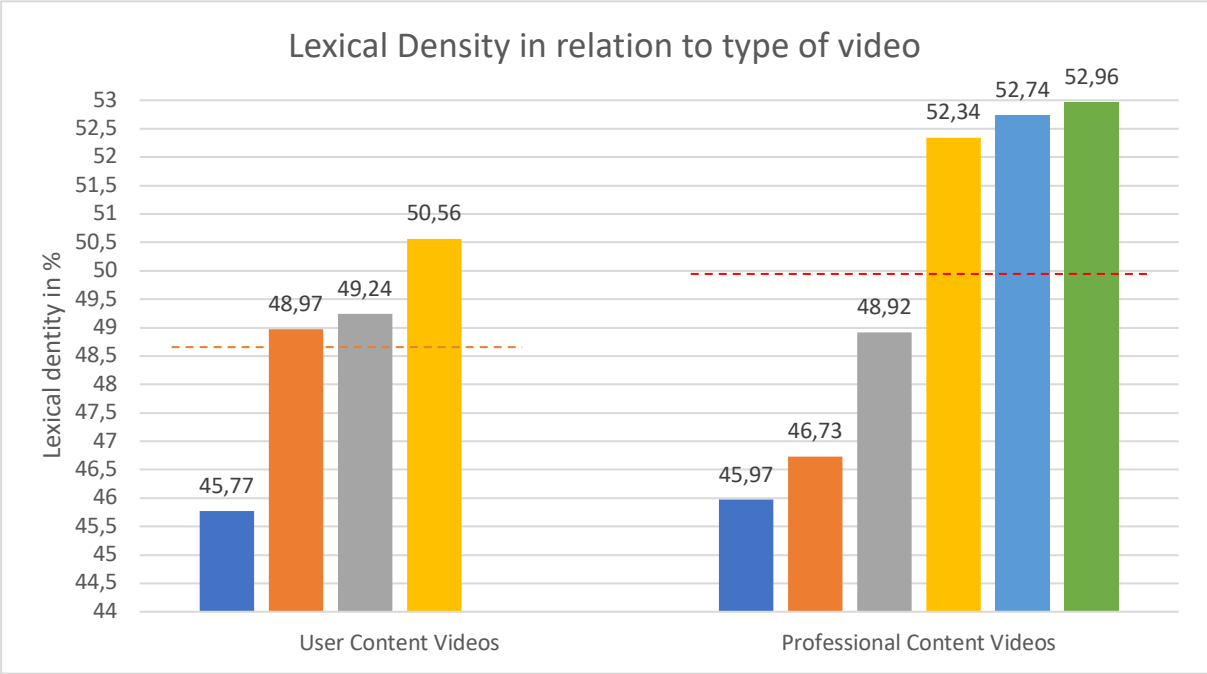

Figure 20: Lexical density in relation to the type of video

As can be seen in the graph, the average of the lexical density (represented by the dotted red line) of professional content videos is more than 1 point higher than the average of lexical density of user content videos[4]. However, what should be underlined is that the difference between the video with the lowest lexical density (45,77 in user content videos and 45,97 in professional content videos) and the highest (50,56 in user content videos and 52,96 in professional content videos) is bigger in the case of professional content videos. This shows that professional videos can also have a low lexical density, but professional videos can have a significantly higher lexical density.

5.3.4 Conclusions regarding Lexical Density

Based on the statistical results, which are consistent with what is visible in the graphs, we can conclude that our data indicates that there is a relation between the lexical density of the texts, their average sentence length and the type of video. This can be motivated by the fact that some videos were made in an informal context and some in a formal context. As stated earlier, occasions in which simpler language is used, lexical density tends to be lower. Why

[4] The average of user content videos is 48,64%, whereas the average of professional content videos is 49,94%.

longer sentences tend to be denser, is harder to determine. A relation between the nativeness of the speaker and Lexical Density was not found.

This means that although our data indicates a link between some of the variables considered, the conclusions have to be taken with a grain of salt because there is not enough data to establish full proof conclusions. The relatively small sample size (10 videos) rather allows for identifying trends in the data. The disadvantage of a small sample size is that it affects the accuracy of the tests, which might make them less reliable (Lane, 2013). Despite this, the observations made in our work are useful for Unbabel, as they allow for identifying patterns found in the data, which can be used as the starting point for future work, namely research exploring with a bigger sample some of the points highlighted here.

5.4 Analyzing Readability

In relation to the readability of the transcriptions of the speech in the video, we decided to make the same type of analysis conducted for lexical density. As with lexical density, the intention was to discover a relationship between *readability*, sentence length, type of video and the nativeness of the speaker in English. To start, the graph below relates the *readability* and the average sentence length.

5.4.1 Readability and Sentence Length

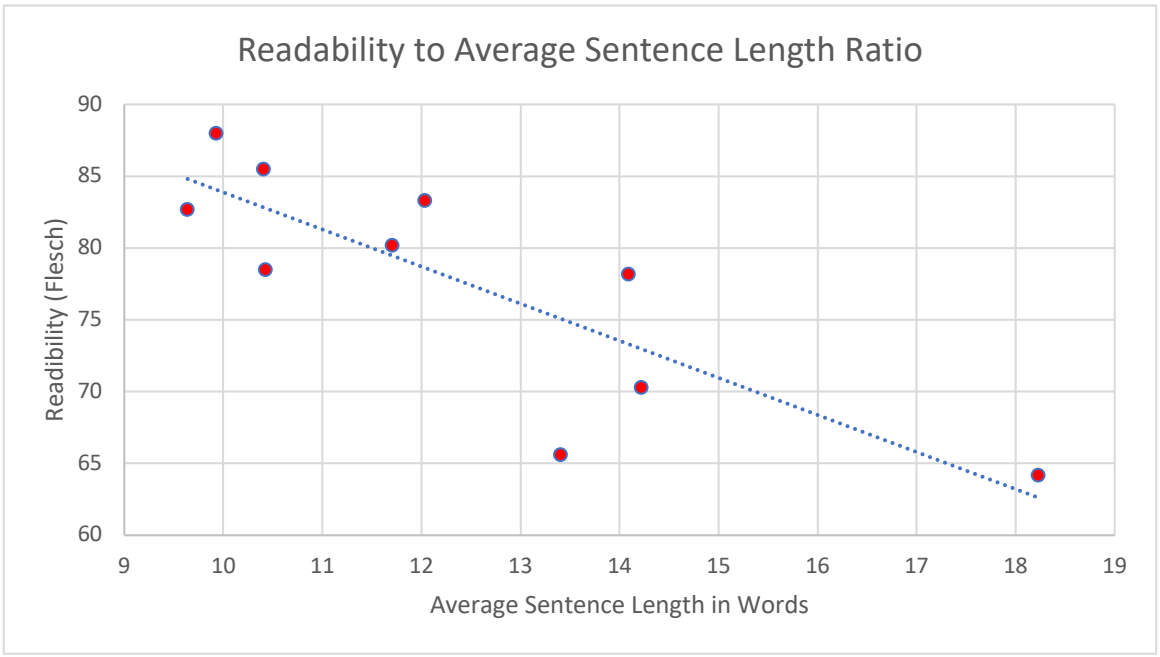

Figure 21: Readability in relation to the average sentence length of the videos

This graph provides a graphical overview of the relationship between the readability and the average sentence length of each video. If we compare “Figure 21” with “Figure 18” it is visible that, in the present case, there is a downward pattern from top left to bottom right. This indicates a negative relation between readability and average sentence length, which makes sense given that Flesch's scoring system works against the grain: the lower the score, the more difficult the text is to read. Like in the lexical density graph, the dots represented in the graph presented above are not far from the scattered line, indicating a relation between the two variables considered: readability and sentence length.

To be able to demonstrate whether there is a significant relation between sentence length and readability we used statistical testing once again. In the present case the null hypothesis (H0) is that there is no relationship between readability and average sentence length, and the alternative hypothesis (H1) is that there is a relationship between both variables. We obtained a p-value of 0.003 (lower than the significance value 0.01 or 0.05) with our data, which demonstrates that that is a statistically significant relation between sentence length and readability in our data.

5.4.2 Readability and nativeness of the speaker in English

Regarding readability, it is also interesting to check whether the fact that the speaker is an English native speaker has an influence on the readability of the text. Once again, we decided to use a graphical representation of the data in the form of a table, presented below.

EN (non-native)			**EN (US)**					**EN (UK)**	
Vid 1	Vid 2	Vid 5	Vid 3	Vid 4	Vid 6	Vid 9	Vid 10	Vid 7	Vid 8
65,6	80,2	83,3	64,2	85,5	78,2	82,7	70,3	78,5	88
Avg: 76,37			**Avg: 76,18**					**Avg: 83,25**	
Total avg (non-native): 76,37 Total avg (native): 78,2 Total avg (both): 77,65									

Figure 22: Readability scores in relation to the nativeness of the speaker

As this table shows, the readability scores for videos with non-native English speakers are in average just over 1 point lower than the scores for the videos with native English speakers. However, the score of the videos with American-English speakers is just slightly lower than the average of videos whose speaker is a non-native English speaker. This would mean that the text of the videos with non-native English speakers are slightly easier to read, because the lower

the score, the more readable the text. The readability results are comparable and consistent with those regarding lexical density because the videos with non-native English speakers were lexically denser (and therefore slightly more complicated) compared to the videos with native English speakers, but slightly lower compared to videos with American-English speakers. Once again, as was the case with lexical density, there is no clear relation found between the readability and the nativeness in English of the speaker in the video. It is interesting, however, that the results regarding readability and lexical density are consistent, indicating a comparable pattern in this respect.

5.4.3 Readability and type of video

To confirm that the results regarding readability and lexical density show similar patterns, readability is also put in relation with the type of video.

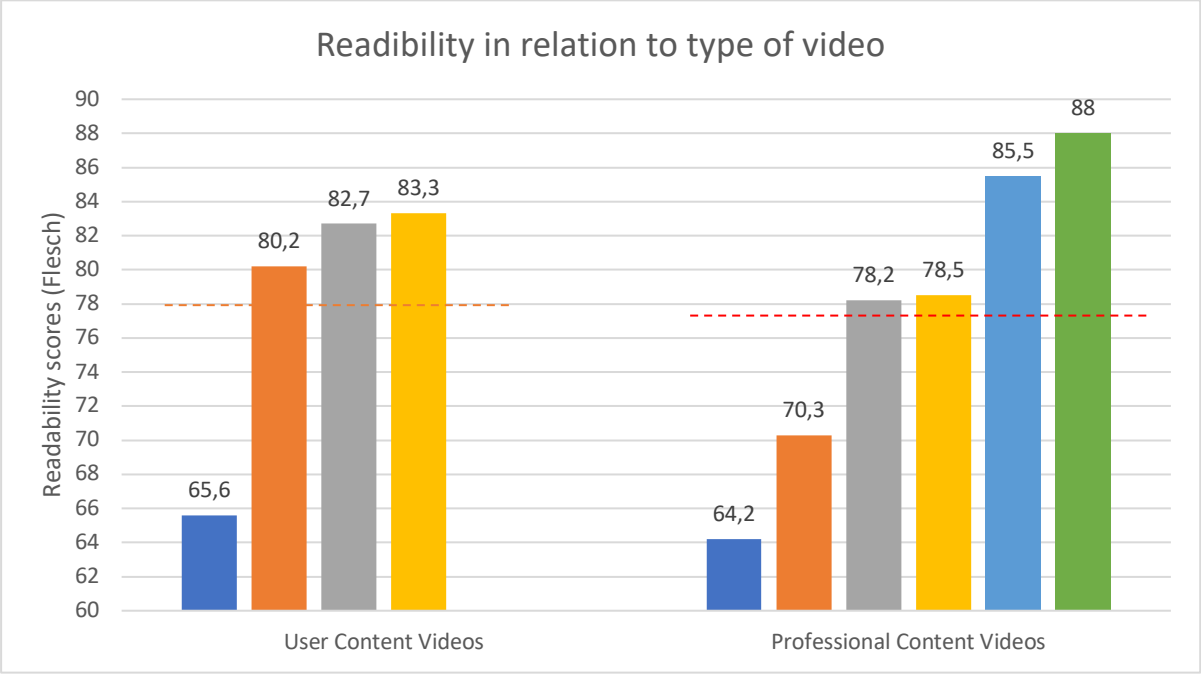

Figure 23: Readability in relation to the type of video

This table shows a different result when compared to the table that shows lexical density in relation to types of video. On difference that might be highlighted is the fact that the averages, in this case, are almost identical. Another difference is, regarding the lexical density of the professional videos, that none of the videos is on the average. All the videos, in this respect, are above or below average, while regarding the readability, for example, two videos are on the

average. It looks like, looking at the professional video, that there is more data dispersion regarding the readability than the lexical density. This is further demonstrated by the values in both tables. In relation to the readability, the lowest value of the professional videos is 24 points lower than the highest value. Next to that, the lowest value of the professional videos is lower than the lowest value in the user content videos, while this is not the case in relation to the lexical density. The final aspect that is perhaps worth mentioning is the fact that the average, in the case of user content videos, is "affected" by one particular video that appears to have a contradictory behavior. All other videos have very similar values of readability, while the value of the video in question has 15 points less. In relation to the lexical density this is also visible to a lesser extent, but not as clearly as in relation to the readability.

5.4.4 Conclusion regarding Readability

The results involving readability show a similar pattern in comparison with the results obtained when lexical density was considered. Like lexical density, there is a statistically significant relation between readability and average sentence length. Also, if we look at the nativeness of the speaker in English and the video type, the patterns observed are also similar, although differences have been observed. These differences may indicate that the relation between lexical density and the other variables is not the same as the relation between readability and other variables, as was made apparent in the graphs presented and commented above. This is in line with other work that has shown that text that is denser is not necessarily more difficult to read (To, Fan, & Thomas, 2013).

Having this said, it seems that readability, due to some differences does not seem to have the same possible links with the average sentence length, type of English spoken and type of video as the lexical density. However, as previously mentioned, the amount of data (sample size) is not enough to draw definite conclusions in this respect, and as showed above, there are other studies that state the opposite.

6. Core analysis

This section describes the "core analysis", in which the errors that were made by the ASR and the MT systems are explicitly described and addressed. By means of tables and examples, we aim at showing a clear picture of the performance of the two systems.

6.1 Problems related to the Dutch Language

First, we decided to highlight errors that appear to be caused by specific characteristics of Dutch. The goal of this section is to describe and explain some issues closely related to specific properties of Dutch, which will probably continue creating problems in machine translation.

6.1.1 Voltooid Tegenwoordige Tijd

One of the first problems is the *Voltooid Tegenwoordige Tijd,* a verb tense that is formed by the verb "hebben (to have) or "zijn (to be)" and a past participle, such as "gevormd ("shaped up" in English)". This verb tense is similar (in terms of structure) to the *Present Perfect* in English, that is formed by the verb "to have" and a past participle. However, despite the similar structure, the *Voltooid Tegenwoordige Tijd* and the *Present Perfect* are associated to different values (Shetter & Ham, 2007). While the *Present Perfect* is used to express an action or situation that started in the past and continues in the present, the *Voltooid Tegenwoordige Tijd* can express a finished action or situation in the past (like the *Simple Past*). A problem caused by this contrast occurred, for example, in line 15 of video 2 in the excel table, illustrated by the example below:

MT	**ASR**
"That kind of **shaped up** the person that I was to become later in my life".	"Dat soort van de persoon **gevormd** die later in mijn leven zou worden".

Example 1: Simple Past vs. Voltooid Tegenwoordige Tijd

Leaving other mistakes occurring in this sentence aside, "shaped up (simple past)" was translated as "gevormd (voltooid tegenwoordige tijd)", without the auxiliary verb "hebben (to have)". To translate "shaped up" we have to use the *Voltooid Tegenwoordige Tijd* in Dutch,

with the auxiliary verb "hebben (to have)" being mandatory. This equivalence between the *Simple Past* in English and the *Voltooid Tegenwoordige Tijd* in Dutch is not correctly handled by the MT system. A different type of problem with the same tense occurred in line 15 in video 3, illustrated by the example below:

MT	**ASR**
"I do not think that I could do this work without having had the experiences that I **had** in special effects".	"Ik denk niet dat ik dit werk zou kunnen doen zonder de ervaringen die ik **had** met speciale effecten".

Example 2: Simple Past vs. Voltooid Tegenwoordige Tijd

In this case, the MT simply translated the verb form "had" in English into the wrong tense in Dutch, while the *Voltooid Tegenwoordige Tijd* should have been used in the translation. These two examples make apparent that the MT system used struggles to select the correct verb tense when translating from English to Dutch. So, not only does it sometimes fail to use the words to construct the tense (like was shown in "example 1"), it also seems difficult for machines to choose the right verb tense in which it should translate. This type of errors were annotated as a *Grammar* errors.

6.1.2 Personal Pronouns with different forms

Another problem that should be highlighted is the fact that in Dutch personal pronouns in the 2nd person (singular) and 3rd person (singular and plural) can have two different forms, which are illustrated in the examples below:

EN	**NL**
"In that restaurant you can get delicious food".	"In dat restaurant kun **je** heerlijk eten".
"You're really crazy!"	"**Jij** bent echt gek!"
"I am a boy, but she is a girl".	"Ik ben een jongen, maar **zij** is een meisje."
"She did not arrive yet".	"**Ze** is er nog niet".

Example 3: Personal Pronouns in Dutch

As "example 3" demonstrates, in Dutch there are different forms for the same person in personal pronouns. In Dutch, these different forms are used when you want to give extra emphasis or when you are making a comparison between two persons (Shetter & Ham, 2007). This is illustrated by the second and third sentence in the example above, in which the personal pronouns have the affix -ij in the end. In some cases, both forms are allowed (like in line 2 of "Example 3"), but there are also cases in which only one form is allowed (like in line 1,3 and 4 of "Example 3"). It seems that the MT system has difficulties in making the distinction between these forms and using them correctly.

6.1.3 Pronouns with a referential function

The MT system has also produced translation errors involving pronouns (both in subject and object positions) when these have a referential function. These errors mainly occur in translating "it". In English, when referring to non-human entities, such as houses, machines etc., "it" is the pronoun mainly used. In Dutch, "it" is often translated as "het", especially in cases where "it" has no referent or when it serves as a subject (Haeseryn, 1984), such as in the sentences: "It is difficult to say (*Het is lastig om te zeggen)*", or "It is a beautiful day (*Het is een prachtige dag*)".

In some cases, "het" can also have a referential function. However, it should first be stated that Dutch has masculine, feminine and neuter words. Neuter words are often combined with the article "het", and these words can also be referred to by the pronoun "het". This is shown in the following example.

EN	**NL**
"The ibis hotel in Amsterdam has many foreign employees. **It** is even run by a Portuguese."	"Het ibis-hotel in Amsterdam heeft veel buitenlandse werknemers. **Het** wordt zelfs gerund door een Portugees".

Example 4: Pronouns with a referential function

It can also serve as a word with a referential function in object position, as seen in this example:

EN	NL
"I gave **it** to him".	"Ik heb **het** hem gegeven".

Example 5: Pronoun with referential function in object position

The problem, however, is that this only applies to neuter words. Words with either a masculine or feminine gender cannot be referred to by "het". In Dutch, depending on the gender and region[5], "hem" or "haar" (object form), or "hij" or "ze / zij" (subject form) are the forms used. This is contrast between English and Dutch is at the basis of several issues in MT outputs. This became evident in an example in lines 13 and 14 of video 6 in the excel table, where this type of error occurs several times in the same sentence:

ASR	MT
"And I'm here to show you how to use **it** and how easy **it** is to use **it**".	"En Ik ben hier om je te laten zien hoe je **het** moet gebruiken en hoe gemakkelijk **het** is om **het** te gebruiken."

Example 6: Pronouns with referential function regarding neuter words

As can be seen in "Example 6", "it" is translated as "het" in all of its three occurrences. In this sentence "it" refers to the washing machine, which can be translated in Dutch as "wasmachine". However, because "wasmachine" is not a neuter word in Dutch, it cannot be referred to by "het". In this case, "it" referring to the washing machine should have translated by "hem" (object form) and "hij" (subject form). Another example of this type of error occurs in line 38 of video 7, demonstrated in the table below:

ASR	MT
"**It's** a little bit dry".	"**Het** is een beetje droog".

Example 7: Pronouns with referential function regarding neuter words

[5] In most parts of the Netherlands, even feminine objects and animals are referred by "hem" (a pronoun that normally refers to masculine words). In southern parts of the Netherlands and Belgium, "haar" is used to refer to feminine objects and animals, but for most of the Dutch people this is unnatural (Haeseryn, 1984).

Here, “it” refers to a burger which is the object of the review in the video and was translated as “het” in Dutch. Once again, “burger” (also “burger” in Dutch) is not a neuter word in Dutch, and therefore cannot be referred to by “het”.

Although we only discuss these two examples, this type of problem is recurrent in our data. We have identified other examples where this phenomenon is at the basis of translation errors in lines 40 and 41, 69-71, and 77 in the same video, but also in video 8 for example.

6.1.4 Postitiewerkwoorden (position verbs)

Another characteristic of Dutch that generates problems in MT outputs involves the so called *positie werkwoorden* (position verbs), also called *locatie werkwoorden* (location verbs), that often cause problems for foreigners learning the language (Lemmens & Slobin, 2007). While several languages only use one or two verbs to indicate a location, in Dutch several verbs are used. The verbs that are considered as position or location verbs are “zitten”, “staan” and “liggen”, which could literally be translated as “sit”, “stand” and “lie (down)” in English. However, these verbs in Dutch are not only used to indicate a movement or position, but also the location. If in English, for example, when you want to say where the car is, you use the verb “to be”, like in the sentence: “The car **is** in the garage”. In this case the verb “to be” helps to indicate the location. However, even though the verb “to be” is often translated as the verb “zijn” in Dutch, in this case that is not possible. It is very unnatural to use the verb “zijn” to indicate where something is unless you don’t know the position of the person or object and the person is in a location in which the position is not clear. This problem will be further explained in the examples below:

EN	**NL**
“The car **is** in the garage.”	“De auto **staat** in de garage”.
“The book **is** on the table”.	“Het boek **ligt** op de tafel”.
“Nick **is** in the car”.	“Nick **zit** in de auto”.

Example 8: Position verbs

The examples show that three different verbs are used in three different situations in Dutch, while in English the verb “to be” is used in all three situations. In these examples, the position

of the object or person are important. In example two, for instance, the book is with its cover on the table, which is why in Dutch the word "liggen" (lie) is used. If the book would be on a shelf, it would be possible to be in a standing position and in that case the verb "staan" (stand) would be used ("Het book **staat** op de plank."). However, as stated before, the verb "zijn", that in many cases is a good translation of the verb "to be", is not commonly used to indicate a location of a person nor object and would not fit in any of the examples given above.

Thus, in general when you want to indicate where something or someone is in Dutch and you know in which position it is, you should use one of the position verbs mentioned above and not the verb "zijn". The MT system however, when translating from English to Dutch, struggles to select the correct position verb. In most of the cases, the MT system translates "to be" as "zijn", as illustrated in the two examples below, taken from line 18 of video 6, and line 56 of video 10:

ASR	**MT**
"Here's your dispenser".	"Hier is je dispenser".
"There are 3 cameras on the roof".	"Zijn er drie camera's op het dak."

Example 9: Position verbs

In both sentences demonstrated in "Example 9" , it is shown that the MT use the verb "zijn" to translate the verb "to be", and in these too, it sounds unnatural, due to the problems explained above. These are two examples, but there are more in which for the translation of the verb "to be", the verb "zijn" is used.

6.1.5 Word order

Another problem that occurred regularly in our data involves the word order used in Dutch. In Dutch, the word order frequently changes in subordinate clauses. This phenomenon is well explained in a paper by Jan Koster, a Dutch linguist, that was published in the Dutch Library for Dutch Literature (Koster, 2002). IThis paper explains that in subordinate clauses in Dutch, the word order changes from subject-verb-object (the word order in main clauses) to subject-object-verb, as illustrated in the following examples:

EN	NL
"Jan buys a second car. **It is cheap**".	"Jan koopt een tweede auto. **Hij is goedkoop**".
"Jan buys a second car, because **it is cheap**."	"Jan koopt een tweede auto, omdat **hij goedkoop is**".

Example 10: Word order in main clauses

In sentence one, there are two independent sentences. The word order is subject-verb-object in both, in Dutch and in English. In sentence two however, although the content if almost the same, the word order changes because there is only one sentence consisting of a main clause and a subordinate clause (connected with the words "omdat" and "because", in Dutch and in English respectively). The example shows that in the Dutch sentence, the order of the main clause stays the same, i.e. SVO, while in the subordinate sentence the words "is" (is) and "goedkoop" (cheap) occur in different positions in the sentence, while in English this does not happen. This is why, according to (Koster, 2002) Dutch is considered to be a "SOV language (Subject-Object-Verb language)".

In this case, however, despite the differences in comparison with English with regard to word order, the MT system uses the right word order in the outputs it produces, thus being able to distinguish main clauses and subordinate clauses, as illustrated in the sentence in lines 44, 45 and 46 of video 6:

ASR	MT
"It actually looks more intimidating if you come into our showroom, because **it doesn't light up**".	"Het ziet er eigenlijk intimiderend uit als je onze showroom binnenkomt, **omdat het niet oplicht**".

Example 11: Word order in subordinate clauses

In this specific case, the MT uses the right word order. If "omdat het niet oplicht" were a main clause, the word order would be different: "Het licht niet op." (it doesn't light up). Generally, in the sentences in our data with a main and subordinate clause, the MT system was able to generate the correct word order.

However, the fact that in main clauses the word order also changes sometimes was at the origin of MT errors identified in our data. Normally the word order in main clauses does not change. However, in sentences with a main clause and subordinate clause where the subordinate clause is at the beginning of the sentence, the word order in the main clause also changes. This can be found in examples taken from line 66 of video 1 and line 72 of video 10, shown in "Example 12".

ASR	**MT**
"And if I'm not wearing my protective gear, **my skin could burn**…"	"En als ik mijn beschermende uitrusting niet draag **mijn huid zou kunnen zijn**"
"Once they have the location of your plate and where you were on that date and time that they scanned your plate, **they can see** where you work, who you associate with, where you pray, where you're going to the doctor."	"Nadat ze de locatie van uw bord hebben en waar je op die datum en tijd was dat ze je bord hebben gescand, **ze kunnen zien** waar je werkt met wie je omgaat, waar je bidt, waar je naar de dokter gaat.

Example 12: Word order with a subordinate clause at the beginning of the sentence

In the examples above, the sentences do not start with the main clause. Both in Dutch and in the English MT output, the main clauses (partly underlined above) start with the subject followed by the verb. Even though Dutch main clauses normally start with the subject and then the verb, in this case this is not the correct word order, due to the fact that the main clause is not at the beginning of the sentence. "Mijn huid zou kunnen zijn" should be "zou mijn huid kunnen zijn" and "ze kunnen zien" should be "kunnen ze zien", where the subject occurs between the auxiliary verb and other verbs. This is a recurrent error in our data. It seems, as stated before, that the MT system can generate the correct word order in Dutch in most cases, even when it is different of the English one, but not always: in main clauses not occurring at the beginning of a sentence, the MT outputs show an incorrect word order.

Another linguistic context in which there is a change in word order in Dutch is *inversion* contexts. Inversion is a reversal of the usual order of phrases, which is especially common in analytic languages, such as Dutch, but also German (Besten & Edmondson, 2002). In Dutch, inversion can occur when a phrase other than the subject is placed at the beginning of the

sentence. Adjuncts, in particular, play an important role in this, because adjuncts, even though normally placed at the end of the sentence, can also be placed in the beginning of the sentence (in front of the subject) (Shetter & Ham, 2007). An example in which this occurs can be found below (taken from line 1 of video 1 in the excel table):

ASR	**MT**
"today **<u>scientists have found</u>** over nine hundred and twenty-five thousand species of insects and there is one man still determined to find more"	"vandaag **<u>hebben wetenschappers</u>** meer dan negenhonderdvijfentwintigduizend soorten insecten gevonden en er is nog een man vast belosten om meer te vinden"

Example 13: Inversion

In this example, in English, the subject comes before the verb, independently of the occurrence of the adjunct "today" in the beginning of the sentence. In Dutch however, due to inversion, the verb, and the subject change position. Therefore "hebben" is placed in front of "wetenschappers". As adjuncts can appear at the beginning of the sentence in Dutch, the word order can sometimes be different than what is generally the case. In this specific example, the MT system generated the correct word order.

Finally, still with regard to word order, it is important to look at the importance of punctuation in relation to word order in Dutch. Periods and commas, for example, do not just separate different sentences from each other, but they also help to distinguish main clauses from subordinate clauses. However, there is often a lack of punctuation in ASR outputs. This generated a significant amount of mistakes in word order in machine translation to Dutch. In our data we found examples where the MT system changed the word order when it should not, as illustrated in the following examples found in line 36 of video 4, line 28 of video 7 and line 56 of video 10.

ASR	MT	Chosen Translation
"Right here you have to do that over 29000 times without making any mistakes and dying **you play** 255 boards".	"Hier moet je dat meer dan 29.000 keer doen zonder fouten te maken en dood te gaan **speel je** 255 borden".	"Hier. Dit moet je meer dan 29.000 keer herhalen, zonder fouten te maken en dood te gaan. **Je speelt** 255 borden".
"…but I thought it'd be an interesting product **to try** no cheese in here".	"… maar dacht dat het een interessant product zou zien om hier geen kaas te **proberen**,"	"… maar dacht dat het een interessant product zou zijn om te **proberen**. Er zit geen kaas in…"
"According to a 2010 study by George Mason University there are three cameras on the roof".	"Volgens een 2010 studie van de George Mason University **zijn er** drie camera's op het dak,"	"Volgens een onderzoek van de George Mason University uit 2010. **Er zitten** drie camera's op het dak".

Example 14: Word order in relation to punctuation

In the first example, due to the lack of a period, it seems that "right here" is in the same sentence, but they should be translated as two different sentences. Because in Dutch, as explained earlier, inversion can occur when there is a phrase other than the subject at the beginning of the sentence, the subject "je" and the verb "spelen" are reversed by the MT system, because it translates the input as a single sentence starting with "hier". The second example shows, as a result of an incorrect identification of the syntactic structure of the source text by the MT system due to the lack of a period, that the verb "proberen" is in a different place in the sentence than it should be. In the third example, too, there is a problem of interpretation due to a missing period, because it seems that the cameras are on the roof according to the study. However, these are two different sentences and due to the lack of punctuation, the MT does not recognize them as such, thus generating an output with a wrong word order. In the MT output the verb "zijn" comes before "er", while "er" should be before the verb.

In general, problems in relation to word order that are caused by contrasts between Dutch and English and by the lack of punctuation, do not cause interpretation problems. However, there are a few exceptions, as can be seen in the examples. Examples of sentences in

which the lack of punctuation causes interpretation problems are discussed in more detail further below.

6.1.6 Note on errors made in relation to Dutch language

After our analysis of the data, we observed that errors apparently related to specific properties of Dutch did not occur in large numbers. These concern about 20 instances. However, it should be noted that this does not necessarily mean that the issues identified above are not an important problem. In fact, as became clear in the previous sections, we were able to identify errors that are systematically produced by the MT system, which allow us to consider that the fact that relatively few errors of this type were identified in our data is simply due to the fact that the contexts and/or phenomena that are problematic to the MT system did not occur often in the source text in our data. It is possible that, for example, when this is not the case, the type of mistakes we observed will occur much more often in MT outputs. That is why it is important that our research will be extended in the future.

6.2 Error analysis in ASR and MT

6.2.1 Global overview

In this section we aim at presenting as much relevant figures and information as possible regarding the performance of ASR and MT systems. In doing so, our main goal is to get a general idea of the errors produced by these systems. The table below shows some general data regarding the errors that were made by both systems.

Number Video	Type of video	Amount of ASR errors	Amount of MT errors	Total	Error to words ratio	Lexical Density	Readibility
1	Professional	56	104	160	0.404	52.96%	65.6
2	Professional	43	53	96	0.482	46.73%	80.2
3	Professional	28	70	98	0.414	52.74%	64.2
4	Professional	74	149	223	0.581	45.97%	85.5
5	Professional	46	64	110	0.337	48.92%	83.3
6	User	41	104	145	0.468	50.56%	78.2
7	User	111	159	204	0.445	49.24%	78.5
8	User	92	162	254	0.589	45.77%	88
9	User	82	158	240	0.552	48.97%	82.7
10	Professional	70	172	242	0.473	52.34%	70.3

Figure 23: General summary of errors made by ASR and MT

To get an idea of whether there is a relation between *Readability* and *Lexical density*, the scores of both methods and the number of errors made in the ASR and MT were included in the table. In previous sections, we were able to identify patterns and trends involving lexical density and readability, when several variables, such the nativeness of the speaker in English, the type of video and sentence length, were considered. Despite the size of our sample, we considered that it would nonetheless be interesting and informative to see if more errors are found in videos with a high *Lexical Density* or low *Readability* score.

The data in the table show that the videos with the highest *error to word ratio* (video 4, 8 and 9) have a lexical density of 48.97% or lower. We can also observe that videos with a relatively high lexical density, such as video 1 and 3, have a low *error to word ratio* compared to the other videos. Given this, it seems that more errors are produced in videos with a low lexical density. However, the video with by far the lowest *error to word ratio*, video 5, has a relatively low lexical density (48.92%). This means that no visible relation between the scores and the number of errors made can be identified in our data.

In terms of readability, we obtain similar results. As for lexical density, videos with a high readability score (and thus easier to read), such as video 4, 8 and 9, have a high *error to word* ratio. Videos 1 and 3 have the lowest scores and have a relatively low *error to word* ratio. However, the video with the lowest *error to word* ratio (5) has a high readability score: the third highest. This means that no visible link between the readability and the number of errors made can be identified.

Finally, the type of video also does not seem to have a clear relation with the number of errors. While some professional content videos, such as videos 1,3 and 5, have a low *error to word* ratio, other videos of this type, such as video 9 and 10, have a relatively high *error to word* ratio. This also applies the other way round. While some user content videos, such as video 2 and 8, have a high *error to word* ratio, other videos of this type, such as video 7, have a relatively low *error to word* ratio.

In addition to an overview of the number of errors made per video, we should also consider the type of errors identified and their severity. Below is an overview showing the number of *minor*, *major,* and *critical* errors per video and which type of error is more common in both the ASR as the MT systems.

Number Video	Amount *Critical* errors	Most common *Critical* errors (ASR – MT)	Amount Major errors	Most common *Major* errors (ASR – MT)	Amount *Minor* errors	Most common *Minor* errors (ASR – MT)
1	47	Named Entity, Overly Literal	38	None, Lexical Selection	83	Punctuation, Punctuation
2	29	Incorrect Word, Untranslated	12	"	65	"
3	14	Incorrect Word, Overly Literal	19	"	54	"
4	57	Incorrect Word, Lexical Selection	32	"	134	"
5	28	Incorrect Word, Lexical Selection	20	"	64	"
6	40	Incorrect Word, Overly Literal	20	"	85	"
7	76	Incorrect Word, Lexical Selection	41	"	168	"
8	93	Incorrect Word, Lexical Selection	31	"	130	"
9	61	Incorrect Word, Lexical Selection	30	"	149	"
10	61	Incorrect Word, Lexical Selection	40	"	141	"
Total	506	Incorrect Word, Lexical Selection	283	"	1073	"

Figure 24: Global overview of severity levels per video

This table shows that *minor* errors occurred most often, followed by *critical* and *major* errors. It also demonstrates that *Punctuation* and *Incorrect Words* are the most common errors produced by the ASR system, while with regard to the MT system *Punctuation* errors, *Lexical selection* errors and *Overly literal* errors are the most common. When video type, readability and lexical density are considered, it is, once again, difficult to say whether there is a relation between the variables. Both user content and professional content videos show *critical*, *major*, and *minor* mistakes, and there is no single type of video which appears to have more or fewer mistakes.

Thus, these overall data does not allow us to establish any clear relation between video types, readability, or lexical density and the amount or type of errors identified. Also, the data presented above do not show the relation between the performance of the ASR system and that of the MT system. We could hypothesize that errors in the ASR output automatically

project an error in the MT. A small overview bellow tends to show to what extent the mistakes that were made by the ASR affect the MT.

Errors in ASR	Errors in MT	Frequency (in sentences)
Yes	Yes	500
No	Yes	134
No	No	73
Yes	No	17

Figure 25: Number of errors ASR and MT at sentence level

In our data there are 724 sentences, 651 of which contain an error (89.92%), whereas 73 do not (11.08%). Focusing on the performance of the ASR, 517 sentences are marked with an error made by the ASR system (71.41%), 500 of which (96.71%) led to an error in the MT output. 634 sentences (87.57%) contain an MT error, 500 of which were also marked with an ASR error (78.86%). In 134 sentences (18.51%) the MT made an error when translating a correct output of the ASR system.

This first table already shows to some extent the influence of the performance of the ASR system on that of the MT system. Below, some simple examples of errors made by the ASR system that also led to an error in the MT output, are shown:

ASR error type	**MT error type**	**Translation**
(1) *"Dan Simon Ø CNN Ø San Leandro Ø California Ø"* (2) *"**he said he** feels it is not only his mission but his responsibility to continue this work."* (3) *"the person that Ø was to become later in my life"* (4) *"okay I'm going to put **the base** on"*	(1) *"Dan Simon Ø CNN Ø San Leandro Ø California Ø"* (2) *"**zei hij dat hij** denkt dat het niet alleen zijn missie is, maar ook zijn verantwoordelijkheid om dit werk voort te zetten."* (3) *"de persoon gevormd die Ø later in mijn leven zou worden,"* (4) *"maar oké, ik ga **de basis** plaatsen"*	(1) *"Dan Simon, CNN, San Leandro, California."* (2) *"**Isaí** vindt dat het niet alleen zijn missie is, maar ook zijn verantwoordelijkheid om dit werk voort te zetten."* (3) *"de persoon gevormd die **ik** later in mijn leven zou gaan worden."* (4) *"Ik ga **dit** opdoen."*
Error type ASR	**Error type MT**	**Description**
1: Punctuation **2:** Named Entity **3:** Missing word **4:** Incorrect word	**1:** Punctuation **2:** Named Entity **3:** Untranslated **4:** Lexical selection	**1:** ASR lacks punctuation, creating a punctuation problem in the MT too. **2:** ASR recorded the name "Isaí" wrongly, leading to an error in MT too. **3:** ASR didn't record the word "I", leaving a part untranslated in the MT. **4:** A misinterpreted word in the ASR leads to a wrong translation in MT.

Example 15: ASR errors that led to an MT error

The first example shows a *Punctuation* error of the ASR system that directly leads to a *Punctuation* error in the MT. This is one of the most common error combinations. Errors of this

kind generally do not present problems for the reader regarding text comprehension but can be bothersome if they happen too frequently. In the *Annotation Guidelines*, punctuation problems are normally rated as *minor* errors. In this work, all *Punctuation* errors have been defined as *minor* as well. However, despite being minor, they can lead to *critical* errors, such as errors that were described in detail in the section focusing on errors related to specific properties of Dutch. In total, in 14 instances a *Punctuation* error led to a *major* error and in 21 instances this type of error led to a *critical* error. This corresponds to 8.27% of the instances in which a punctuation error in the ASR output led to an error in MT.

The second example is an example in which the error produced by the ASR system is responsible for an interpretation problem. "Isaí" has been recognized and transcribed by the ASR system as "he said he", which lead to a total incorrect translation, which can therefore mislead viewers of the video. These types of errors happened frequently during the analysis and are more problematic, as they can significantly mislead the viewer.

The third error is also a critical one, because the ASR system was not able to recognize a word at all, which was therefore not translated by the MT system. An untranslated segment usually causes problems in text comprehension.

In the fourth example, there are not missing words in the ASR output, but a word has been incorrectly recognized and transcribed. The word "this" has been transcribed as "the base", which is why the MT system also translated "the base" as "de basis" in Dutch where it should have been translated as "dit" (this).

What is interesting is that errors where a word is missing or is incorrect occurred more often in videos whose speaker is either not an English native speaker or is a British-English native speaker. For example, in the videos whose speakers have British-English as their mother tongue or are not native speakers of English (4 videos) 58 incorrect word and 9 missing word errors were made, while 35 incorrect word errors and 11 missing word errors were made in the rest of the videos (6 videos). Also, by default, these types of errors are problematic and considered as *critical errors*.

Next to these errors, there were several other types of errors that were made by the ASR system and that either lead or did not lead to MT errors. To get a better idea of what kind of errors are caused by both systems, below the types of errors produced by each system are described in detail, with examples and tables, allowing us to observe how often a certain type of error has been made.

6.2.2 ASR errors

6.2.2.1 Punctuation errors

The examples that were shown in the table above are also generally the most common errors produced by the ASR system. One of the most common type of error found in ASR outputs is the *Punctuation* error. We found 481 *Punctuation* errors produced by the ASR system. This means that, in our data, given an ASR error (there were 517 in total) there is a big chance that it is a *Punctuation* error. Out of the 481 *Punctuation* errors produced by the ASR system, 58 did not lead to an error in MT (12.06%), meaning that the other 423 *Punctuation* errors identified led to an error in MT (87.94%). In fact, considering the 367 *Punctuation* errors produced by the MT system, only 3 were produced by the MT itself. The remaining 364 had an ASR error at its origin, meaning that 99.18% of the *Punctuation* errors found in MT outputs were caused by the performance of the ASR. As stated before, in general a *Punctuation* error also causes a *Punctuation* error in the output of the MT system: 364 cases in which an ASR *Punctuation* error leads to an MT *Punctuation* error, corresponding to 86.05% of the instances in which an ASR *Punctuation* error leads to an error in MT.

However, as showed earlier, ASR *Punctuation* errors can also lead to other types of error in the translation, which can be *major* or *critical* errors, and thereby mislead viewers when reading the text or the subtitles. A clear example of this follows below:

ASR	**MT**	**Translation**
"I think my microscopes Ø my camera's Ø vials and anything else I might need to document this Expedition"	*"Ik denk dat mijn microscopen,* ***de flacons*** *Ø* ***van mijn camera*** *en al het andere dat ik nodig heb om deze expeditie te documenteren,"*	*"Ik neem mijn microscopen,* ***flacons, camera's*** *en al het andere dat ik nodig heb om dit onderzoek vast te leggen."*
ASR error type	**MT error type**	**Description**
Punctuation	Lexical selection	Missing punctuation leads to interpretation error

Example 16: Punctuation errors

In the example is shown that due to the lack of commas, the text is misinterpreted by the MT system. While the words “cameras” and “vials” should be translated separately, as independent items in an enumeration, in the MT output the vials are part of the camera. In fact, the MT output is literally “the vials of the camera” (“de flacons van mijn camera”). What is interesting is that despite the lack of punctuation in the ASR output, the MT system still manages to introduce commas in the output it produces. However, the MT system is not able to solve all the *Punctuation* errors produced by the ASR system. For example, in addition to fact that some commas continue to be missing in the MT output, no periods are present.
Another example is presented below, in which a *Punctuation* error leads to a *Lexical Selection* error, which is considered a *critical* error, as it can mislead the reader.

ASR	**MT**	**Translation**
“no Ø explain it to me”	*“leg het me **niet** uit”*	*“leg het me uit”*

Example 17: Punctuation errors leading to a critical error

In this example, it appears that, due to a missing comma, the MT system interprets “no” as negation of the main verb “explain”, which would be canonically be expressed by “don't”, which causes the translation produced to be incorrect and misleading, as the MT output states exactly the opposite that is being said by the speaker. This error has also been defined as a critical *Lexical selection* error.

This means that serious errors like these occur more than once in our data: there are 16 cases in total (3.78% of ASR errors that lead to a MT error) in which an ASR *Punctuation error* leads to a *Lexical Selection error* in the MT output, 11 of which were considered *critical* (68.75%), 2 *major* (12.50%) and 3 *minor* (18.75%). At first 3.78% may seem a relatively small percentage, but if we look at the number of ASR *Punctuation* errors that caused an error other than a punctuation error (69), this percentage is not that small (23.19%), corresponding to almost a fourth of all MT non-punctuation errors caused by the ASR. Additionally, as most of these errors were considered to be critical, thus being prone to cause significant problems for understanding the video content, they should therefore not be ignored.

Another type of MT error recurrently caused by ASR *Punctuation* errors is the *Grammar* error. There are 32 of such cases in our data, which corresponds to 46.38% of MT non-punctuation errors caused by an ASR *Punctuation* error, meaning that the *Grammar error*

is the most common "not *Punctuation* error" that is caused by a *Punctuation* error in the ASR. According to the *Annotation Guidelines*, *Grammar* errors should be considered as *minor* or *major*, although some of these errors have been considered *critical* too. Of the 32 *Grammar* errors identified in MT outputs that were caused by ASR punctuation errors, 15 were defined as *minor*, 12 as *major* and 5 as *critical*. In the table below some examples are presented:

ASR	**MT**	**Translation**
(1) *"with a low carbon footprint Ø I always manage to bike Ø hike Ø swim or rap Ø"* (2) *"but guys Ø let me know in the comments below if you tried this Ø"* (3) *"so we will see Ø guys Ø"*	(1) *"met een lage CO2-voetfdruk Ø **lukt het** altijd om te fietsen, wandelen, zwemmen of rap,* (2) *"maar jongens **laten het me** weten in de reacties hieronder als je dit hebt geprobeerd,"* (3) *"dus we **zullen jongens zien**,*	(1) *"met een lage CO2-voetafdruk. **Het lukt** me altijd om te fietsen, wandelen, zwemmen, of raften.* (2) "*Jongens, **laat me** weten in de reacties hieronder of je hem hebt geprobeerd."* (3) *"Dus we **zullen zien, jongens**."*
ASR error type	**MT error type**	**Description**
Punctuation	Grammar	Punctuation errors cause a grammatical error in MT.

Example 18: Punctuation errors leading to grammar errors

As explained earlier, punctuation errors can cause grammatical errors when translating into Dutch. This has to do with the grammar rules of the language itself. In the first example, which illustrates this type of error, the lack of a period to distinguish the two sentences in the ASR output, leads the MT system to translate it as a single sentence, starting with an adjunct phrase "with a low carbon footprint", followed by the main clause "I always manage to bike, hike, swim or raft". Earlier in this chapter I showed that in Dutch the word order can change if the main clause is not at the beginning of the sentence. This is what happened in the MT output. The words "lukt" and "het" changed positions, although this should not happen, as both words introduce a new sentence, starting with a main clause. This error creates some confusion,

because in the output produced by the MT system it seems that it is because of the footprint that the speaker manages to hike, swim or raft. However, in the video these were two separate sentences, and thus the aforementioned semantic relation should not be in the translation.

In line 2 and 3 of "Example 18", *Grammar* errors cause problems too. In the second example, the absence of a comma causes the MT system to misinterpret the text. Without the comma, it seems that "guys" is the subject, thus letting the object "me" know something. However, "guys" is just a vocative, a term used by the speaker to address the audience, whom the speaker kindly asks to inform him about something. A comma, that is lacking in the ASR output and thus in the MT output, would contribute to avoid confusion. In English, the lacking comma does not have a much bigger impact than that, as in English the third person plural and the second-person of the imperative of the verb "let" have the same form ("let"). In Dutch, however, they are different ("laten" and "laat", respectively). This fact creates additional problems in understanding the subtitles for the Dutch viewer: in the MT output there seem to be "jongens" (boys) that are going to inform the speaker about something.

In the third example there is a similar problem, even though the Dutch translation is probably even a bit more confusing. Because of the missing comma, the word order changes and therefore the meaning. Once again due to a missing comma, "guys" is considered to be the object that is seen by the subject "we". In Dutch, unlike in English, the object is placed between the auxiliary verb and the main verb. Because the MT system considers that "guys" is the object, the word order changes.

Of course, as shown before, most MT *Grammar* errors caused by ASR *Punctuation* errors are *minor*. However, 12 *major* and 5 *critical* errors of this type is not something to ignore, which justifies the interest of the data description presented above.

6.2.2.2 Incorrect word errors

Another type of error that occurred regularly in the ASR outputs is the *Incorrect word* error. In 14.64% of the sentences with an ASR error (642), this was the type of error identified (94 cases). This type of error is often significantly problematic because, as described in the *Annotation Guidelines*, not only does it almost always lead to an error in the MT output, but also, and more importantly, it can seriously affect the comprehension of the subtitles. According to the guidelines, an *Incorrect word* error can be considered of any severity level, depending on its impact. In our data, all *incorrect word* errors were considered *critical*. The MT errors caused by ASR *Incorrect word* errors were all *Lexical selection* errors. The following table shows examples this type of error, thus making apparent how serious these can be:

ASR	**MT**	**Translation**
(1) *"I couldn't wait to share it with the rest of the **work**"* (2) *"it moved **the** Donkey Kong"* (3) *"Britain **The Source just that are just looked** off basically **just hope that America"***	(1) *"Ik kon niet wachten om het met de rest van het **werk** te delen,"* (2) *"het **bracht de** Donkey Kong **in beweging**,* (3) "*Britain **The Source er gewoon zo uitziet**, **alleen maar hopen dat Amerika***	(1) *"Ik kon niet wachten om het met de rest van de **wereld** te delen."* (2) *"Daarna **ben** ik **overgegaan op** Donkey Kong."* (3) *"**De saus, ik lik het er net van af... smaakt eigenlijk gewoon naar mayonaise**."*
ASR error type	**MT error type**	**Description**
Incorrect word	Lexical selection/Overly literal	Type of error in which the ASR recognizes a word or sequence of words wrongly. This creates a wrong MT.

Example 19: Incorrect word errors

The first sentence in the table is a simple example in which one word has been mistranscribed by the ASR system (The word "work" should have been transcribed as "world"). Although the two words have a similar orthographic form, they have a completely different meaning, both in English and in Dutch.

The second example also shows that even a small spelling mistake can lead to a serious translation error. The ASR system transcribed the word "the" instead of "to", a switch that can cause a lot of trouble in translating into Dutch. Additionally, regarding this example in particular, the verb "moved" in English can be used in multiple contexts, while Dutch does not have a word that would be a good equivalent for the word "moved" in these multiple contexts. For instance, "moved the" and "moved to" in English are very similar in form, which is why

they might not be that confusing if the rest of the context is available. In Dutch, however, different words should be used to accurately translate “moved to” and “moved the”. Therefore, in the chosen translation, the verb “in beweging brengen” was replaced by the verb “overgaan op”.

The third sentence in the table is an example in which the ASR completely mistranscribed the entire sentence. What the speaker is saying is “the sauce ... oh I just licked off”. Words or verbs like “sauce” and “source”, and “licked off” and “looked off” are still similar in terms of spelling, but the rest of the words that the ASR transcribed are not close at all to what was really said. This led to a complete mistranslation.

The examples above show how *Incorrect Word* errors can significantly affect the understanding of the text. This does not mean, however, that all of them do. *Incorrect Words* do not always cause *major* or *critical* errors. This is shown in the following example.

ASR	**MT**	**Translation**
*“and I just **tuck** this in my trolley based on **and** Friends review of it”*	*“en ik **stop** dit in mijn trolley op basis van **en** Beoordeling voor vrienden,*	*“Ik **heb** deze net in mijn koffer **gestopt** vanwege een recensie van **een** vriendin.”*

Example 20: Incorrect word errors that do not cause a major error

In this example, the ASR system transcribed the words “tucked” and “an” as “tuck” and “and”. The words do not differ much from each other in terms of orthographic form and, at least “tuck” and “tucked”, are forms of the same verb. The word “an”, although far from being similar in meaning to “and”, does not cause a lot of confusion hindering the understanding of the text. In Dutch it is also obvious, due to the context, that “een” was meant, rather than “en”. The verb “stoppen”, due to the ASR error, was translated in the present tense instead of the past tense, but also in this part it is possible to understand what is meant.

However, *Incorrect Word* errors that do not cause significant problems are rare. From all ASR *Incorrect Word* errors that led to an error in the MT output (89), 88 were considered *critical*. This is 15.55% of all ASR errors that led to an error in the MT output. This means that most of this type of error has an important impact in the MT output and can really affect the viewer's understanding of the text.

6.2.2.3 Missing word errors

Another type of error made by the ASR system that caused problems in the performance of the MT system is the *Missing Word* error. According to the *Annotation Guidelines*, these are errors in which the ASR did not transcribe a word that was in the audio. Missing words automatically result in a word or several words to be untranslated by the MT system, whereas there is also a case in which it led to a different translation. These errors in MT, therefore, are classified as *Untranslated* errors or *Lexical selection* errors. In the following table we show some examples of sentences in which ASR *Missing word* errors led to MT *Untranslated* errors or *Lexical Selection* errors.

ASR	**MT**	**Translation**
(1) *"this Ø pre-30 stumbles upon a new pool of water among the vegetation"* (2) *"every time I've gone one stage I have to make sure my audience Ø with me"* (3) *"after I have time to pour over the footage and consult Ø very smart man and the subject*	(1) *"deze Ø pre-30 struikelt over een nieuwe plas water tussen de vegetatie."* (2) *"elke keer als ik het podium op ga moet ik ervoor zorgen dat mijn publiek bij mij Ø"* (3) *"na Ik heb tijd om de beelden over te gieten en **de** zeer slimme man en het onderwerp te raadplegen"*	(1) *"Dit **jaar** stuitte Isaí op een nieuwe plas water temidden van planten."* (2) *"Elke keer als ik het podium op ga, moet ik ervoor zorgen dat mijn publiek met mij **meeleeft**."* (3) *"Nadat ik de tijd heb gehad om de beelden te bestuderen en **een** hele goede kenner van het onderwerp te raadplegen."*
ASR error type	**MT error type**	**Description**
Missing word	Untranslated, Lexical selection	Errors in which the ASR system was not able to transcribe everything that the speaker was saying, automatically causing a part of what the speaker says to remain untranslated or be wrongly translated by the MT system.

Example 21: Missing word errors

In the first example, in addition to the mistranscribed name “pre-30” (discussed in another section), the word “year” is not included by the ASR output. As a result, the word “jaar” has been omitted from the MT output, although it should be in the translation (see “translation” column). This results in important information being missing in the translation, but also in establishing incorrect syntactic relations: the word “this” is a demonstrative, and due to the absence of the word “year”, it is put in a syntactic relation with “pre-30”, resulting in an incorrect translation. Because the word “jaar” is missing, the wrong demonstrative was used in the Dutch translation. As explained earlier, “dit” or “dat” are the correct demonstrative pronouns in Dutch to refer to neuter words. The MT system used “deze”, which is used for gendered words, while “dat” should have been used.

In the second example, the ASR failed to transcribe the verb "are". Since the missing element is a verb, it can pose major difficulties for understanding the subtitles, as verbs are known to be the words that form the basis of a sentence (Shetter & Ham, 2007) and therefore sentences without a verb lack important information. In the absence of the verb in our example, the context is not sufficient for the viewer to guess which verb should be there.

In the first two examples, the *Missing Word* errors lead to serious translation errors, but this is not always the case, as illustrated in our third example. In this example the ASR system failed to include the article "a", leaving it also untranslated in the MT. What is interesting is that, despite the *Missing Word* error, the MT system added it in the translation. It seems that if an article is missing, the MT system adds one anyway. In this specific case, the MT system chose the definite article “de”. Given that there are not other sentences in which a similar error occurred in our data, we cannot determine whether the MT system always selects for a definite article when an article is missing in the source text or not, or whether if it systematically adds an article at all. In the example in the table, as mentioned, the MT system chose the article “de”, which is not the correct one, because the speaker uses the indefinite article “a”, which should therefore be translated as “een”. Nevertheless, this does not pose any major problems for understanding the subtitles. The articles “a” and “een” indicate the definiteness of a noun phrase but are not words that significantly contribute to the semantic content of a sentence. At the most, the error might lead the speaker to think that a specific “man” is being referred to, but since there is no other reference to such a “man” in the rest of the video, the context clarifies that there no specific person is being referred to. The ASR system made 20 *Missing Word* errors, which correspond to 3.12% of the errors that were made by this system. It is the fourth most common type of ASR error, after *Punctuation*, *Incorrect Word* and *Named Entity* errors. 13 of the ASR *Missing Word* errors led to an error in MT (65%). This means that there were 7

instances in which a *Missing Word* error did not cause an error at translation level. This is an apparently high number, because there were cases in which *crutch words*[6] such as "um" were not transcribed. However, as *crutch words* are often omitted in the translation, their absence was not identified as a MT error. Of the 13 *Missing Word* errors that caused an error in MT, only the example in the table, which is a *Lexical Selection* error, can be considered an error that does not cause many problems for the understanding of the text. All the other errors can be considered *critical* errors, as they significantly affect the understanding of the subtitles. This means that 60% of ASR *Missing Word* errors, and 92.31% of the ASR *Missing word* errors that caused an error in MT, are *critical*.

6.2.2.4 Named Entity errors

Another type of error that has occurred about as often as the *Missing Word* error, is the *Named Entity* error. As explained in the *Annotation guidelines*, this is an error in which names of persons, places, locations, or entities in the source text do not match those of the target text. This error occurred 24 times in the ASR outputs, corresponding to 3.74% of all ASR errors. All of them led to an error in MT. In fact, considering all ASR errors that led to an error in MT, but were not *Punctuation* errors, 16.78% were *Named Entity* errors. The table below lists some typical *Named Entity* errors:

[6] A crutch word is a word that becomes a filler in conversation, or is used for verbal emphasis, without any meaning to an utterance (Doll, 2012)

ASR	MT	Translation
(1) *"**John George** is a famous chef and he owns a three-star Michelin restaurant in New-York which I used the work at* (2) *"**my cats like Harbour** found what he says is an egregious violation of privacy* (3) *"it is a 24-hour lip color I'm not sure if that's focusing from **Ellie**"*	(1) *"**John George** is een beroemde chef-kok en hij heeft een driesterren Michelin-restaurant in New York waar ik werkte"* (2) *"**mijn katten zoals Harbour** ontdekten dat wat hij zegt een flagrante schending van de privacy is"* (3) *"het is een 24-uurs lipkleur ik weet niet of dat de focus is van **Ellie***	(1) *"**Jean-Georges** is een beroemde chef-kok en hij heeft een driesterren Michelin-restaurant in New York, waar ik heb gewerkt."* (2) *"**Mike Katz-Lacabe** vindt dat het naar zijn mening een grote schending van de privacy is."* (3) *"Het is een 24 uur langhoudende lippenstift, ik weet niet zeker of het soft focus is, van de **Aldi**."*
ASR error type	**MT error type**	**Description**
Named Entity	Named Entity	Name of person, place, location, or entity is mistranscribed by the ASR system, causing an error in MT.

Example 22: Named entity errors

In the first example the ASR system transcribed "Jean-Georges" as "John George", a *Named Entity* error that does not have a major impact on the output of the system, as it does not affect the general semantic content of the text. However, it can be misleading for people very interested in "gastronomy", as probably there is not a "John George" who owns a restaurant in New York, and if a "John George" actually does exist, the audience may confuse both names. According to the *Annotation Guidelines*, *Named Entity* errors are, by default, marked as *major* errors, but in this example, it is difficult to say how severe the error is. It depends whether the viewer is familiar with famous chefs or not. Taking this into account, the error was considered *critical*.

The second example also describes an error that caused a lot of trouble. The ASR misunderstood the name "Mike-Katz Lacabe" for "my cats like Harbour", causing a translation error. Normally, errors that are caused by ASR *Named Entity* errors and lead to a mistake in MT, are also annotated as MT *Named Entity* errors, but the error could also easily be annotated as a *Lexical Selection* or *Overly literal* error, because it was not even translated as a name. Since a name was not transcribed and, therefore, machine translated as a name at all, but as a phrase, its impact on the understanding of the subtitles by the viewer is major.

In the third example, as in the first, the ASR transcribed a name, but opted for the wrong one: "Ellie" should have been transcribed as "Aldi". This can be very confusing as "Ellie" is a person's name and "Aldi" is the name of a supermarket. In the specific context of this examples, it can make people think that the lipstick belongs to a particular person or that the lipstick is from a particular brand. Due to this, this error could also be evaluated as a *critical* error.

Of the 24 *Named Entity* errors identified in our data, 17 led to a misleading translation, which corresponds to 70.83% of all cases. This means that 19.17% of the cases were not critical. In these cases, only some spelling mistakes were made which did not affect text comprehension. An example of a case like this can be found in video 2, lines 26 and 27, where the ASR system transcribed "Charlie", while it should be "Charley".

As discussed in detail in (Gannay, et al., 2020), ASR systems are very sensitive to *Named Entity* errors and even the most modern ASR systems fail to accurately interpret more than 70% of the names, although this depends on the language used in the source text. 24 errors out of 642 ASR errors seems a relatively low figure, but it does not mean that this is an insignificant type of error. As mentioned earlier, some errors are so specific that they are dependent on the source text. If there are no names to transcribe in the source text, then there are not *Named Entity* errors in the target text. That is why it is important, once again, that in future research more data is gathered. If there had been more data with more names, the percentage of named entity errors would possibly be higher, as modern ASR systems continue to struggle to recognize names.

6.2.2.5 Extraneous word

The last type of ASR error made in this dataset and discussed here is the *Extraneous word* error. This is a type of error in which the ASR system adds a word that is not present in the audio. As it adds information to the input of the MT system, this type of error leads the MT system to translate segments of text that do not have to be translated. As described in the *Annotation Guidelines*, this type of error is likely to confuse the viewer and cause

misunderstanding of the subtitles. In total, 11 *Extraneous Word* errors were made, corresponding to 1.73% of the ASR errors, 9 of which led to an error in MT, which corresponds to 6.29% of the MT errors that were caused by the ASR but not by a *Punctuation* error. The table below shows some typical *Extraneous Word* errors:

ASR	**MT**	**Translation**
(1) *"really **them** the only flavor I'm going from the Patty is General vegetable and coriander"* (2) *"**Britain** The Source just that are looked off basically just hope that America"* (3) *"**And** so I basically didn't even know what this was until I checked the menu on their website…"*	(1) *"echt de enige smaak die ik ga van de Patty is General groente en koriander"* (2) *"**Britain** The Source er gewoon zo uitziet, alleen maar hopen dat Amerika"* (3) *"**en** zo ik wist eigenlijk niet eens wat dit was totdat ik het menu op hun website controleerde,…"*	(1) *"Het enige wat ik over het algemeen proef van de burger, zijn de groenten en koriander."* (2) *"De saus, ik lik het er net af… smaakt eigenlijk gewoon naar mayonaise."* (3) *"Ik wist eigenlijk niet eens wat dit was, totdat ik het menu op hun website zag…"*
ASR error type	**MT error type**	**Description**
Extraneous word	Untranslated, Addition	Added word(s) by the ASR lead to the inclusion of information in the translation that should not be there, leading to confusion for the viewer.

Example 23: Extraneous word errors

In the first example, the word "them", which is not present in the audio, has been added by the ASR system. What is interesting is that the MT system did not include this word in the translation, which is the correct choice in this case. In most cases, depending on the context, the word "them" would be translated as "ze" or "zij" in Dutch, but in the first example it is not

present in the MT output. So, in this example where, despite the inclusion of an additional word that is not present in the audio in the transcription produced by the ASR system, the MT makes the right choice by not including it in the translation.

But the second example presented above shows that this type of error can cause a lot of confusion. In this example, the ASR included the word "Britain" in the transcription, leading the MT system to keep this word in the translation. The fact that the word is also left untranslated ("Britain" in Dutch is "Britannië) by the MT system, will probably be even more mindboggling to the viewer. In the final translation, besides the correction of other errors occurring in this sentence, this word was omitted.

In the third example, the word "and" is added by the ASR system. This is, however, an example where the unnecessarily added word does not seriously affect the understanding of subtitles. This is because, in this case, the subtitles would be understandable both with or without the word "and". The word "and ("en") is a transitional word[7] that expresses addition, but in this case, there is nothing being added, which makes it actually more confusing than removing it.

Of all errors, the first and third examples are the least common, meaning that most *Extraneous Word* errors caused severe problems in the output, like in example 2. Besides the second example that was shown in the table, there were 7 other cases in which an *Extraneous Word* caused a *critical* error, meaning that 88.89% of the *Extraneous word* errors that lead to an error in MT lead to a critical error.

6.2.3 MT errors

This section focuses on the errors identified in MT outputs, especially the ones that were not caused by the ASR system, as these were already highlighted and discussed in previous sections. As shown in the summary in a previous section, in our data there are 634 sentences in which a MT error was identified. In total, there were 1220 errors made by the MT system, i.e. an average of 1.64 errors per line and approximately one mistake per each three words (counting all types of mistake). In the table below is shown which type of MT errors are most common.

[7] Transitional words refer to words that assist in the logical flow of ideas and connect sentences and paragraphs (Jackson, 2005)

Type of error	Instances	Percent
Lexical selection	416	34.10%
Punctuation	367	30.08%
Grammar	186	15.25%
Overly Literal	104	8.52%
Untranslated	38	3.11%
Should not be translated	36	2.95%
Named Entity	29	2.38%
Mistranslated Term	18	1.48%
Addition	18	1.48%
Locale Conventions	6	0.49%
Capitalization	5	0.41%

Figure 26: Overview type of errors made by MT

The table shows that *Lexical Selection* errors, *Punctuation* errors, *Grammar* errors, and *Overly Literal* errors are the most common MT errors. The other types of errors occur less often in comparison with these. However, as some of the type of errors that do not occur often, tend to be *critical* (as explained in the section on ASR errors), these should not be ignored. This table also shows that *Punctuation* errors from the MT perspective occur often. It is even the second most often type of error in MT. This makes sense, because in the previous section it became clear that almost all ASR *Punctuation* errors also lead to an error in MT, and *Punctuation* errors are the most common in the ASR.

6.2.3.1 Lexical selection errors

Although *Punctuation* errors are very frequent, the most common error in MT outputs is the *Lexical Selection* error. As can be seen in the previous table, about one third of the total number of MT errors made are *Lexical Selection* errors. 111 of the 416 *Lexical Selection* errors were caused by the ASR (26.68%), which means that 305 were generated by the MT itself (73.32%). The table below shows some examples of *Lexical Selection* errors:

ASR	MT	Translation
(1) *"I make artificial patients to allow real doctors to practice procedures to improve health outcomes for kids"* (2) *"some of them are really low-tech like an arm that you can inject or practice line placement"* (3) *"it's also extremely important to e PSI D that he's able to get to each new location with a low carbon footprint* (4) *"you can always ask me if you have any questions*	(1) *"Ik maak **kunstmatige patiënten** zodat echte artsen **procedures** kunnen toepassen om de **gezondheidsresultaten voor** kinderen te verbeteren."* (2) *"sommige zijn **echt low-tech** zoals een arm dat u lijnplaatsing kunt injecteren of oefenen"* (3) *"het is ook uiterst belangrijk voor PSI D dat hij in staat is om naar elke nieuwe locatie te **komen** met een lage CO2-voetafdruk* (4) *"je kunt me altijd vragen **of je vragen hebt**"*	(1) *"Ik maak **poppen**, zodat echte artsen **werkwijzen** kunnen toepassen om de **gezondheidstoestand van** kinderen te verbeteren."* (2) *"Sommigen zijn **heel eenvoudig**, zoals een arm om te injecteren of om het inbrengen van een infuus te oefenen."* (3) *"Het is ook uiterst belangrijk voor Isaí dat hij in staat is om naar elke nieuwe locatie te **gaan** met een lage CO2-voetafdruk."* (4) *"Je kunt me altijd iets vragen **als dat nodig is**."*
ASR error type	**MT error type**	**Description**
None	Lexical selection	Translation contains words that do not accurately convey the meaning of the original word in the source text.

Example 24: Lexical selection errors

As described in the table, following the *Annotation Guidelines*, *Lexical Selection* errors are errors in which the translation does not accurately convey the meaning of the source text. What often happens is that a word used in the translation is an equivalence of the source word, but it does not fit the context of the source text or is simply not accurate in that situation. This is an issue that occurs often in translation, and the first example illustrates this clearly, with four

Lexical Selection errors in a single sentence. All words used in the translation are Dutch words, but they are used in different contexts or situations. The last word is particularly interesting to analyse, because it is a preposition that is used in different contexts in the two languages. In many contexts "voor" is an accurate translation for the English word "for", although not in the specific context of our example. In this context, "van" (closer to "of" in English) should be used. In the analysis of our data, we identified several cases in which the most usually correct translation of a particular preposition was not suitable for the specific context of the source text.

The second example illustrates a case in which the equivalent chosen by the MT system is a possible equivalent of the source text in the relevant context, but is not often used by Dutch speakers. The word "low-tech" exists and is used also in Dutch, but mostly by people that work in technical areas, or, have at least some knowledge of these areas. Because the word is only used by these types of people and rarely, its use can be confusing to the general public. Therefore, we chose for a more general translation as "eenvoudig" (simple). These kinds of less accurate translation choices were found frequently in our data and the harder question is whether such a *Lexical Selection* error is *minor*, *major,* or *critical*. It is an error in which the translation conveys the correct meaning, but it can generate text comprehension problems because the equivalent present in the translation is not commonly used.

The third example shows another problem that is also found sometimes in our data. The MT used "komen" for translating "get to". Although, this is an accurate equivalent in many contexts, in this example "gaan" should have been used. This also happened in a few occurrences of the verb "to come", where "gaan", which is normally used to translate the verb "to go", would be the equivalent to be used. In fact, if we consider dictionary definitions in English and in Dutch, the verbs "to go", "gaan", "to come" and "komen" are associated to the same definition (Merriam-Webster Incorporated, 1999) (Van Dale Uitgevers, 2015): the verbs "to go" and "gaan" are described as "moving to a certain direction away from the speaker", while the verbs "to come" and "komen" are described as "reaching a certain destination (towards the speakers)". However, although these verbs make the same general semantic contribution, they do not have exactly the same distribution in the two languages, as made apparent by our examples.

In the fourth example, there is a combination of *Lexical Selection* errors. First, we have an issue related to the word "if", in English, which can introduce a conditional clause as well as an indirect question ("He asked **if** I had left with you") (HarperCollins, 2011). In Dutch however, different words have to be used for each of these linguistic contexts, which means that the word "if" cannot always be translated by the same word in Dutch. To

introduce a conditional clause, “als” is the equivalent to be used, whereas “of” should be used to introduce an indirect question. Chosing the right translation of “if” remains a problem for the MT system, as several cases in our data make apparent. Additionally, there is also a problem in our example related to the translation of the phrase “if you have any questions”. The issue in this case is related to the fact that “vragen” is a plural noun that is also an equivalent of the verb “to ask”. Therefore, it is used two times “vragen” in a single sentence, to translate the noun “questions” and the verb “to ask”. Both expressions are totally common, also in this context, but despite being allowed to be used in the same sentence, the sentence does not sound natural due to repetition of the same words, which is why we proposed another translation. However, there is not an accurate error type to describe a case such as this where the translator edits the MT output, simply to avoid a repetition of the same word forms. An example such as this one shows that either the annotation guidelines or the error typology itself could be improved to cover this kind of situation.

Most *Lexical Selection* errors found in our data are similar to those described in the first example: sentences in which the equivalent used by the MT system is not the correct equivalent in a particular context. These are classified as *major* errors as they could create problems for the reader, but not problems that make it impossible to understand the text. However, there are *Lexical Selection* errors that were classified as *critical*. The section dedicated to ASR errors already showed that many critical *Lexical Selection* errors were caused by ASR *Incorrect Word* errors (88), but additionally, there are also 51 errors not caused by the ASR that were classified as *critical Lexical Selection* errors, meaning that the MT system itself is responsible for introducing critical translation errors in the outputs. Besides the aforementioned 51 critical errors, from the *Lexical Selection* errors that were not caused by the ASR (305), 232 errors were considered *major*, and 27 *minor*. This means that most of the *Lexical Selection* errors (even including the ones caused by the ASR), create severe comprehension problems.

6.2.3.2 Punctuation errors

The *Punctuation* errors caused by the ASR were addressed at an earlier stage in this report. It was shown then that most of the MT *Punctuation* errors were caused by the ASR. The table below shows some examples in which the ASR was not responsible for *Punctuation* errors in the MT output.

ASR	**MT**	**Translation**
(1) *"Messiah Nakamoto who's the Godfather of video games he crowned me video game player of the year for that and all the achievements"* (2) *"people ask me what do you do when you're not cooking"* (3) *"I'm a chef in Bangkok and I have a restaurant in my home called the table"*	(1) *"Messiah Nakamoto, de Ø peetvader van videogames Ø, hij kroonde me Ø game player van de eeuw Ø daarvoor en alle prestaties"* (2) *"Mensen vragen me Ø wat je doet als je niet aan het koken bent Ø* (3) *"ik ben een kok in Bangkok en ik heb mijn restaurant in mijn huis genaamd Ø the table Ø"*	(1) *"Masaya Nakamura, de 'vader van videogames' heeft me daarvoor en voor alle vorige prestaties tot 'gamer van de eeuw' gekroond."* (2) *"Mensen vragen me: "Wat doe je als je niet aan het koken bent?"* (3) *"Ik ben een chef-kok uit Bangkok en run een restaurant in mijn huis, genaamd 'The Table'."*
ASR error type	**MT error type**	**Description**
None	Punctuation	Punctuation error that was not caused by the ASR but is still present in the translation.

Example 25: Punctuation errors caused by MT

In the first example, a problem occurs because the subject of the sentence consists of a nickname and a name. In Dutch it is common to put nicknames, such as "the Godfather of video games", between quotation marks. This is also the case with titles such as "video game player of the year". This error is not counted as an ASR error, as I cannot say, as a non-English native, whether in English quotation marks are used for nicknames and titles.

In the second example, there is a *Punctuation* error related to citations. In Dutch a colon (:) is used to introduce a citation introduced by an expression such as "he said" (Shetter & Ham, 2007), while in English a comma is used in this context (Warriner, 1981). This means

that after the word “me” we should have a colon in the Dutch translation, as shown in the “translation” column.

The third example is an example in which a proper name occurs. In Dutch, when a proper name (in this case the name of a restaurant) occurs, especially if it is a name in a different language, it appears between quotation marks, like in “The Table”. The fact that terms and names in another language are placed between quotation marks in Dutch is the reason why an example such as the one we are addressing here is not an error of the ASR, but an error in the translation introduced by the MT system.

The examples in the table are the only *Punctuation* errors not caused by the ASR found in our data and they were all considered *minor* errors. This shows that it is not very likely that the MT introduces *Punctuation* errors by itself and, when it happens, the impact to the comprehension of the text tends to be low. Because these are errors that are not caused by the ASR, it is also possible that more errors similar in nature to the ones discussed above will occur if more data is considered. The *Punctuation* errors that cause significant impact in translation quality, as shown in the section on ASR errors, are ASR errors that lead to other types of errors in the translation, and not ASR *Punctuation* errors that cause MT *Punctuation* errors or *Punctuation* errors caused by the MT itself.

6.2.3.3 Grammar errors

Grammar errors are the third most common MT error found in our data. Of the 186 grammatical errors that occurred, 33 (17.74%) were caused by the ASR, while 153 (82.26%) were not caused by the ASR, showing that this is an error mainly caused by the MT system itself. In the section dedicated to ASR errors, it was shown that *Grammar* errors caused by the ASR could lead to some serious problems. For example, it became clear that some grammatical errors were made due to a combination of a lack of punctuation and contrasts in English and Dutch grammar, and that a part of these errors were considered *critical*. These were addressed in the section on ASR errors. In the table, we show examples of *Grammar* errors that were caused by the MT system itself.

ASR	MT	Translation
(1) *"I make artificial patients to allow real doctors to practice procedures to improve health outcomes for kids"* (2) *"while I was working for special effects I got the opportunity to build dinosaurs for the Jurassic Park theme parks and magical creatures for Harry Potter and all kinds of wonderful things to bring that experience to life for people"* (3) *"chickpeas so you kind of expect that"* (4) *"he crowned me video game player of the century for that and all the achievements*	(1) *"Ik maak kunstmatige patiënten Ø zodat echte artsen procedures kunnen toepassen om de gezondheidsresultaten voor kinderen te verbeteren"* (2) *"terwijl ik werkte met speciale effecten die ik kreeg de mogelijkheid om dinosaurussen te bouwen voor de Jurassic Park-pretparken **en** magische wezens voor Harry Potter en allerlei prachtige dingen om die ervaring tot leven te brengen voor mensen.* (3) *"**kikkererwten is**, dus je verwacht dat,"* (4) *"hij kroonde me game player van de eeuw daarvoor en alle prestaties"*	(1) *"Ik maak poppen, zodat echte artsen werkwijzen kunnen toepassen om de gezondheidstoestand van kinderen te verbeteren."* (2) *"Toen ik met speciale effecten werkte, kreeg ik de mogelijkheid om dinosaurussen te maken voor de Jurassic Park-pretparken, magische wezens voor Harry Potter en allerlei prachtige dingen om die ervaring voor mensen tot leven te brengen."* (3) *"**Het zijn kikkererwten**, dus dat verwacht je wel".* (4) *"...heeft me daarvoor en voor alle vorige prestaties tot 'gamer van de eeuw' gekroond".*
ASR error type	**MT error type**	**Description**
None	Grammar	Error made by the MT system that is related to the grammatical structure of the target language.

Example 26: Grammar errors caused by MT

In the first example the problem is linked to the verb "to allow". In Dutch there is not an equivalent of the verb "to allow" in English that is adequate in every context. In many cases,

the verb “toestaan” is used, but in this context it is not accurate. A nice option is a construction with the transition of causality “zodat” and the verb “kunnen”, as proposed by the MT system. This construction is comparable with the construction “so (they) can” in English. In Dutch, however, as “zodat” is a transition that links clauses, it is mandatory to put a comma before it.

In the second example, there is an unnecessary repetition of the conjunction “en (and)”. In Dutch, when there is an enumeration, different items are separated from each other with a comma, instead of using iteratively using the word “en (and)”. Therefore, in the translation we used a comma, rather than the word “en” repeatedly. In English this is also recommended and common (McCaskill, 1998), meaning that the error is probably caused by the quality of the source text produced by the speaker. However, even if the writing style, or “speaking style” in this case, does not of high quality, it is often considered as the responsibility of the translator to solve such problems in writing style (Nida, 1984). Therefore, the error was annotated as an error that was not caused by the ASR, but by the MT system.

The first and second examples, show errors considered to be *minor*. The third example, however, involved an error consiederd to be *major*. This is because in this case the speaker “swallows” words: instead of saying “they are Chickpeas” the speaker only said “Chickpeas”. As there is a verb missing in the source sentence, the MT system has trouble translating it. Although the MT system add the verb “zijn”, it uses it in the wrong tense. Also in Dutch it is necessary to add “het” (they), which is not included in the MT output. As, despite the issues described above, it is still possible to understand the content to some extent, this was considered to be a *major* error and not a *critical* one.

The fourth example contains a *Grammar* error that was annotated as *critical*. In this example, the word order is so confusing that it does not allow the viewer to understand what is conveyed by the speaker, which led this error to be considered *critical*. Critical *Grammar* errors t caused by the MT system only happened 3 times, but this shows that the MT system can be at the origin of critical *Grammar* errors, which cause severe problems in relation to the text comprehension.

The errors described in the table are the type of grammatical errors caused by the MT that occur most frequently. Most of them (120) were considered to be *minor*. However, there were also 30 *major* errors and 3 *critical*, meaning that 21.57% of the *Grammar* errors made by MT was at least defined as *major*. Considering also the fact that 17 ASR *Grammar* errors were defined at least as *major* errors, these are numbers that should be taken into account and addressed in the future, as *Grammar* errors, whether caused by the ASR or not, tend to create major problems in translations.

6.2.3.4 Overly literal errors

According to the *Annotation Guidelines*, "an issue is an *Overly Literal* error when the source text is translated in a very literal way, which may result in problems of interpretation". In the data considered in this work, 104 errors were marked as *Overly Literal*, which correspond to 8.52% of all the MT errors, none of which were caused by the ASR. The table below shows some *Overly Literal* errors.

ASR	**MT**	**Translation**
(1) *"It moved the Donkey Kong but* ***the greatest level of competition at the moment*** *was Pac-Man"* (2) *"****I'm just like a vessel*** *but through me they can feel the emotions* (3) *"****so it tells me here*** *estimated time remaining of 59 minutes"*	(1) *"het bracht de Donkey Kong in beweging, maar* ***het grootste niveau van concurrentie op dit moment*** *was Pac-Man"* (2) *"****Ik ben net een schip,*** *maar door mij kunnen ze de emoties voelen* (3) *"****dus het vertelt me hier*** *de geschatte resterende tijd van 59 minuten"*	(1) *"Daarna ben ik overgegaan op Donkey Kong, maar* ***het meest competitieve spel*** *toendertijd was Pac-Man."* (2) *"****Ik draag ze als het ware over****. Door mij kunnen ze de emoties voelen."* (3) *"****Hier staat*** *de resterende geschatte tijd van 59 minuten* ***aangegeven****."*
ASR error type	**MT error type**	**Description**
None	Overly literal	Source text was translated by the MT system too literally.

Example 27: Overly literal errors

The examples in the table are all very similar: they are all examples in which a certain expression was translated too literally. This means that this type of error often involves a sequence of words, not just one or two. This is the case, for example, in an expression, in which a large part of a sentence is often mistranslated as a whole and when doing so several words are chosen incorrectly. Despite this, the mistranslated sequence of words is counted as a single error (an *Overly Literal* error).

In the first example, the sequence “the greatest level of competition at the moment” was literally translated by the MT system, using the most common translation of each individual word occurring in it. In Dutch, instead of the literal equivalent of “the greatest level of competition” (het hoogste niveau van concurrentie), one should use the translation “het meest competieve spel toendertijd was Pac-Man”, literally corresponding to “the most competitive game at the moment was Pac-Man” in English.

In the second example, the phrase “I'm just like a vessel” was problematic to the MT system. A vessel can be “a watercraft bigger than a rowboat”, but also “a tube or canal in which body fluid is contained and conveyed or circulated” (Merriam-Webster Incorporated, 1999). In the context at stake is not one completely clear what the speaker means, but he most likely wants to say that he sees himself as someone that serves as a carrier of emotions. The context clarifies that the speaker thinks that his audience can feel the emotions through him. In Dutch, the term “vessel” is not used in any common expression in either of the senses considered above. Despite this, the MT system translated the expression literally, using the translation for “vessel”, in the sense of a boat. For this phrase, the translation “Ik draag ze als het ware over” fits better, which in English could be literally translated as “I transfer them in a way”, referring to the emotions that he transmits.

The third example shows a literal translation of the expression “it tells me here”. In English, it seems common to use the verb “to tell” to indicate something that is written somewhere. In Dutch however, the verb “staan” (often translated as “to stand” in English) is the verb usually used for expressions that indicate something that is written. For these reasons, we chose the expression “hier staat aangegeven”, which literally corresponds to “here stands indicated” in English. In this expression it is not natural to use articles and personal pronouns, which explains why the words “het” and “me” were omitted from the translation.

The other *Overly Literal* errors present in our data are all similar in nature to the errors included in the table. As mentioned before, *Overly Literal* errors are different from *Lexical selection* errors because *Overly Literal* errors tend to involve sequences of words rather than isolated words. Additionally, in *Lexical selection* errors the expression generated by the MT system is used in the target language, although in a different context. Translations of expressions that are labeled as *Overly Literal* are often not used at all in the target language. This explains the fact that these types of errors are caused mainly by the MT system and not by the ASR system. Another aspect specifically related to this type of error is that this is a type of error that often involves specific cases, such as expressions or idioms. This means that the amount of *Overly Literal* errors closely depends on the source text. If the source text contains

a lot of idioms, the probability that more *Overly Literal* errors will occur is higher. Also, *Overly Literal* errors tend to create a lot of problems in relation to text comprehension: as mentioned, only one *Overly Literal* error was not considered *critical*.

6.2.3.5 Untranslated errors

Untranslated errors occur when content that should have been translated is left untranslated. According to the *Annotation Guidelines*, by default these errors are *critical*, because they are likely to affect the way the audience parses and understands the video. *Untranslated* errors occur about just as frequently as *Shouldn't be translated* errors in our data. In total, 38 *Untranslated* errors were made, corresponding to 3.11% of all MT errors. 18 of them were caused by the ASR, corresponding to 47.37% of the *Untranslated* errors. Most of them, as briefly explained in the section focusing on ASR errors, were caused by *Missing Word* errors, but there were also cases that were caused by *Capitalization* and *Named Entity* errors. As these errors were not specifically addressed in the ASR errors section, some examples are included in the table below, in parallel with the *Untranslated* errors that were caused by the MT system itself:

ASR	MT	Translation
(1) *"we identify quite a lot with country music because of the message that is in it family love heartbreak"* (2) *"really them only flavor I'm going from the **Patty** is **General** vegetable and coriander"* (3) *"you can't miss a **DOT** a prize a blue man you can't die once Ø can't lose focus*	(1) *"identificeren we ons vrij veel met countrymuziek vanwege de boodschap die erin zit familie Ø liefdesverdriet"* (2) *"echt de enige smaak ik ga van de **Patty** is **General** groente en koriander"* (3) *"je kunt een **DOT** een prijs niet missen een blauw man die je niet kunt sterven als je eenmaal niet kunt sterven, je kunt de focus niet verliezen"*	(1) *"We identificeren ons regelmatig met countrymuziek vanwege de boodschap die het brengt. Familie, **liefde**, liefdesverdriet."* (2) *"Het enige wat ik over het **algemeen** proef van de **burger**, zijn de groenten en koriander."* (3) *"Je mag geen **bolletje**, krachtpil of blauw spookje missen. Je mag niet één keer doodgaan. Je mag de focus niet verliezen."*
ASR error type	**MT error type**	**Description**
None, Named Entity, Missing word	Untranslated	The MT system left a part of a text that should have been translated untranslated.

Example 28: Untranslated errors

The first example was not caused by the ASR. The words "family" and "lovebreak" were translated as "family" and "liefdesverdriet", which are good choices, but the word "love", for some reason, is left untranslated when it should have been translated as "liefde", a word cannot be omitted, as it has a different meaning.

The second error shown in our table is an error caused by the ASR. This tool capitalized the words "Patty" and "General", perhaps because the ASR interpreted them as names. In fact, "General" is sometimes used to indicate a name of a war commander, in which case it should be capitalized in English. However, this is not the case in our example, where "general" occurs in the sense of "common" or "widespread". In addition, "Patty" is sometimes used for the name

of a person, but in this case it refers to a burger. The ASR *Capitalization* error can explain why the MT system left these elements untranslated, as they were treated as names. However, in this context, these words correspond to a noun and adjective that are indispensable for understanding the text.

In the third example, we have a similar error, also caused by the ASR. This time the word "dot" was transcribed in capitals. In the game world, DOT is an acronym for "damage over time", "referring to acts that slowly causes damage to characters, such as poison" (Dictionary.com, 1995). However, in our example, the word "dot" refers to the small dots that Pac-Man must eat in the game to get rid of the ghosts and go to the next level, and thus should not have been transcribed in capitals. Due to the ASR error, this word is left untranslated in the MT output, which seriously affects the viewer's understanding of the text.

Of the ASR errors that cause an *Untranslated* error in the translation, the examples given in the table (2 and 3) are the only ones that are not caused by a *Missing Word* error, but rather by *Capitalization* and *Named Entity* errors. It seems that sometimes the ASR sees words as names or the other way round. In this case, it is also a bit unclear if these should be addressed as *Capitalization* errors or *Named Entity* errors, because the capitalization of the words is a result of the fact that the ASR thinks that they correspond to a name. All errors caused by the MT itself are similar to the type of error described in the first example. For inexplicable reasons, text that should be translated was omitted in the MT output. In all of our examples this type of error seriously affects the viewer's text comprehension. All these errors create confusion and can make the text to be interpreted in a different or wrong way, and were therefore considered *critical*. Examples 1 and 2 show that an *Untranslated* error, which has always a significant impact in translation quality, can be caused by an error that is normally, according to the *Annotation Guidelines*, considered to be *minor*.

6.2.3.6 Should not be translated errors

Should not be translated errors are issues in which text that should have been left untranslated, was translated. In our data we found 36 of these errors, 4 of which were caused by the ASR. According to the *Annotation Guidelines*, this type of errors are by default *critical*. In the table below, some typical *Should not be translated* errors are shown:

ASR	MT	Translation
(1) *"That **kind of like** shaped up the person that I was to become later in my life"* (2) *"hi I'm Rachel from **parenting** Central and today I'm going to do something a little bit different"* (3) "*There was nothing saying **congratulations** there was nothing saying **the end game over** nothing"*	(1) *"dat **soort van** de persoon gevormd die later in mijn leven zou worden"* (2) *"hoi ik ben Rachel van **ouderschap** Central en vandaag ga ik iets anders doen"* (3) *"maar er was niets te zeggen **gefeliciteerd**, er was niets **dat het eindspel over niets** zei,*	(1) *"Dat heeft de persoon gevormd die ik later in mijn leven zou gaan worden."* (2) *"Hoi, ik ben Rachel van '**Parenting** Central' en vandaag ga ik iets anders doen."* (3) *"Er stond geen: '**Congratulations'**. Er stond geen: '**The end'. 'Game Over'**. Niets."*
ASR error type	**MT error type**	**Description**
None	Should not be translated	A part of the source text that should not have been translated was translated.

Example 29: Should not be translated errors

In the first example, the problem is the usage of "like" as a *crutch word*, i.e. "a word that becomes a filler in conversation, or is used for verbal emphasis, without any meaning to an utterance" (Doll, 2012). In Dutch, but also possibly in English, *crutch words* are often omitted in transcriptions and subtitles, precisely because they do not add any meaning to an utterance and tend to affect the readability of the subtitles (Hoek & Sonépouse, 2012). The expression "kind of like", especially "like", which is shown in the example, is a word often used in spoken English (McWhorter, 2016). The MT system, however, translated this expression literally, thus causing confusion while adding little or nothing at all to the utterance. What should be noted here is that this example is not a typical example of a *Should not be translated* error. Normally, like in line 2 and 3 of "Example 29", a typical *Should not be translated* error is an error in which a word or sequence of words in the source text should be remained the same in the target text.

In this case, the problem is more whether *crutch words* should take place in written texts (such as subtitles) or not. Being so, this example could be addressed as another type of error, such as a register error. However, also the typical register errors do not adequately correspond with types of errors such as crutch words, which is why we addressed the error as a *Should not be translated* error.

In the second example, the ASR is responsible for a translation that was not necessary. "Parenting Central" is the name of the company of the speaker, but the ASR system did not transcribe this expression as a name, as it did not use a capital letter in "parenting", although it did so in "Central". Because "parenting" has no capital letter, and therefore is not recognized as a name, the MT system translated it when it should have kept it unchanged.

In the third example, the speaker is referring to a part of a text that appears on the screen of a gaming machine in a game hall, such as the text "the end"or "game-over" if you lose or die. Because most gaming machines are imported from America, and texts are usually left untranslated in the Netherlands, it is unnecessary and even confusing to translate these texts. However, the MT system translated them.

The *Annotation Guidelines* refer to *Should not be translated* errors as errors occurring when "text that should have been left untranslated was translated (e.g., brands, foreign words, etc.)". As mentioned, by default, these errors are critical. In the second and third examples presented above, this is the case, as both errors generate a very confusing output and affect the viewer's text comprehension. However, in the first example we can see an error which may be bothersome, but does not really lead to a completely erroneous interpretation of the text. From the 36 *Should not be translated* errors, only 7 were considered *minor*. The rest of them are like the error shown in the first example. Given these observations, there should be a revision of the *Annotation Guidelines*, with *Should not be translated* errors not being considered *critical* by default, or with the creation of a new error typology to account for crutch words.

6.2.3.7 Mistranslated term errors

According to the *Annotation Guidelines*, "an issue is a *mistranslated term* error when the translation of a certain term is not the preferred/appropriate one". This type of error is the specialized domain version of the *Lexical Selection* error. General verbs, for example, are considered as *Lexical Selection* errors as they can be used in any context. Considering its specificities, the frequence of occurrence of this type of error too depends to a great extent on the type of source text. If only a few terms that are used in a specific area in the source text, *Mistranslated Term* errors will therefore not be frequent. In total, we found 18 *Mistranslated*

Term errors in our data, which corresponds to 1.48% of MT errors, with only one being caused by the ASR. In the table below, some typical *Mistranslated Term* errors are shown:

ASR	**MT**	**Translation**
(1) *"he tracks to the **Patagonian ice field** searching for new species that have been hiding for 200 million years"* (2) *"I'm a chef in Bangkok and I have a restaurant in my home called the table"* (3) *"here's your dispenser where you'll put your detergent fabric softener and liquid chlorine bleach"*	(1) *"hij volgt **de Patagonische ijsveld** op zoek naar nieuwe insectensoorten die zich al meer dan 200 miljoen jaar verbergen"* (2) *"ik ben een **kok** in Bangkok en ik heb mijn restaurant in mijn huis genaamd The Table."* (3) *"hier is je **dispenser** waar je je **wasmiddel**, **verzachter** en **vloeistof** in doet"*	(1) *"Hij gaat naar de **Patagonische ijsvelden** om nieuwe insectensoort te ontdekken, die zich al meer dan 200 miljoen jaar schuilhouden."* (2) *"Ik ben een **chef-kok** uit Bangkok en ik run een restaurant in mijn huis, genaamd 'The Table'."* (3) *"Hier zit de **wasmiddellade** waar je je **wasmiddel**, **wasverzachter** en **bleekwater** in doet."*
ASR error type	**MT error type**	**Description**
Punctuation (?), None	Mistranslated term	A specific word that relates to a particular area or field was translated wrongly.

Example 30: Mistranslated term errors

The first example illustrates one of the few errors that were considered *minor*. The problem is that the term "Patagonian Icefields" was used in singular form in the source text. The "Patagonian Icefields" are often divided into the "Northern Icefield" and "Southern Icefield" (MDPI, 2019). This indicates that if you refer to one of the icefields, you are allowed to use the singular form, but not if you refer to both, which is apparently the case in our example because the speaker does not mention a specific icefield. So, as has been mentioned in other occasions in this report, it seems that the error is in fact created by the speaker. This error causes the term to be translated in singular form in the target text, when it should be used in the plural.

This error can be considered *minor* as it has a very limited impact in terms of the understanding the subtitles and it does not change the meaning of the original message.

In the second example, the term "chef", referring to "a professional cook, typically the chief cook in a restaurant or hotel" (Oxford Languages, 2009), should be translated in Dutch by the term "chef-kok", which is defined as "een kok die leiding geeft" (a chef that is in charge) (Van Dale Uitgevers, 2015). The term "kok", that was used by the MT system is not equivalent to the term "chef-kok", designating a professional in this area in a lower rank. A "chef-kok" is always the cook that is in charge, while a "kok" is a cook that works for someone else. As the speaker is the owner of her own restaurant, thus being in charge, the right term should be "chef-kok" and not "kok". This error is from more serious since there is a difference in meaning between both terms. Viewers of the video will understand the speaker's profession, but important information may be lost in relation to the rank she holds, if the term "kok" is used instead of "chef-kok".

In the third example, multiple terms of the same type are used in a single sentence. The terms refer to cleaning products and washing machines. The first term to be highlighted is the term "dispenser". The MT system chose to leave "dispenser" untranslated. This is understandable, as the word "dispenser" exists in Dutch. However, in English, the term "dispenser" refers to "a container or device for holding and dispensing small amounts" (Farlex Inc., 2003), which can also be a "container" in the washing machine. In Dutch, a "dispenser" is mainly a small box that is used to store pills (Van Dale Uitgevers, 2015). Next to that, even for a pillbox, the term "pillendoos" is preferred to "dispenser". In the context of washing machines, the term "dispenser" is not used in Dutch at all, which means that the MT system mistranslated the term. The MT system also had problems translating the term "liquid chlorine bleach", which was translated as "vloeistof". "Vloeistof", however, is a general word that roughly corresponds to the word "liquid" in English, which can be used in a lot of contexts. This means that the word used in the translation only covers the semantic contribution of "liquid" and not of "liquid chlorine liquid bleach". The correct equivalent for "chlorine liquid bleach" is "bleekwater". In parallel to these incorrectly translated terms, in this example there are terms which were (partly) well translated by the MT system. For example, "verzachter" is not officially the term that is used to translate "softener" ("wasverzachter" is), but it is likely that Dutch viewers will understand what is meant. "wasmiddel" is also an adequate equivalent for the term "detergent".

The examples in the table are representative of *Mistranslated Term* errors occurring in our data. These are errors where a term used in a specific area is incorrectly translated. As mentioned, *Mistranslated term* errors, unlike *Lexical selection* errors, are very specific and only

occur in specialized discourse. This is probably one of the reasons why there is a significantly lower amount of *Mistranslated Term* errors than *Lexical selection* errors. If no specific terms are present in the source text, then no *Mistranslated term* errors will be made. The examples in the table showed that this type of error can generate a lot of confusion for the viewer, often hindering the correct interpretation of the subtitles. This is also why the *Annotation Guidelines* describe this type of error as *major* or *critical* by default. The error in the first example is the only *Mistranslated Term* error considered *minor*. All other *Mistranslated Term* errors found in our data were *critical*.

6.2.3.8 Addition errors

The *Addition* error is the last type of error that is discussed in detail in the present work. An *Addition* error occurs when there is text in the target text that is not present in the source text. According to the *Annotation Guidelines*, these are likely to affect the parsing and understanding of the subtitles by the viewer. In total, 18 *Addition* errors were found, corresponding to 1.48% of MT errors. In the table below, some typical *Addition* errors are shown.

ASR	MT	Translation
(1) *"this here is an infrared camera police tell us that technology is only beneficial agains Crime suspects"* (2) *"why is that confusing"* (3) *"I'm going to use my camera as a mirror so like bear with me when I get super close"*	(1) *"hier is een infraroodcamera politie **die** ons vertelt dat technologie alleen gunstig is voor verdachten van misdaad,"* (2) *"waarom is dat **is** verwarrend,"* (3) *"Ik ga mijn camera gebruiken als een spiegel, dus draag **als ik bij me kom** als ik heel dichtbij kom,"*	(1) *"Dit is een infraroodcamera. De politie meent dat de technologie alleen gunstig is voor verdachten bij misdrijven,"* (2) *"Waarom is dat verwarrend?"* (3) *"Ik ga mijn camera als een spiegel gebruiken, dus vergeef me als ik heel dichtbij kom."*
ASR error type	**MT error type**	**Description**
Any	Addition	MT added text that is not present in the source text. Most likely this affects the interpretation of the subtitles by the viewers.

Example 31: Addition errors

As mentioned in the ASR section, *Addition* errors can be caused by ASR errors. In 8 of the 18 *Addition* errors this was the case, as illustrated in the first example. In English relative pronouns can be omitted when they are the object of the clause. Due to the lack of punctuation in the ASR output, the MT system considers that a relative pronoun was omitted in the source text, when it is not the case, and introduces it in the translation, since it is not possible to omit relative pronouns in Dutch. 6 of the 8 *Addition* errors that were caused by the ASR involve a problem regarding relative pronouns.

The errors in the second and third example were caused by the MT itself. There are 10 *Addition* errors in our data that were caused by the MT itself. In the second example, the word "is" was added, thus creating serious problems for understanding the text in the video, and therefore being considered a *critical* error.

In the third example the sequence “als ik bij me kom” was added to the translation in Dutch, while nothing is said by the speaker which could explain its inclusion in the target text. Thus it was classified as an *Addition* error that was not caused by the ASR.

All of the 8 *Addition* errors that were caused by the ASR were due to a lack of punctuation. Most of them involve the addition of a relative pronoun, as mentioned earlier. In most of the cases where an *Addition* error occurred, it had a great impact on the meaning of the sentence, often leading to a misinterpretation by the viewer. Thus, these errors were considered *critical*. Only one *Addition* error was considered *minor*, in which the word added did not have any impact on the meaning of the rest of the sentence.

7. Findings and Discussion

In this section, the main findings that can be derived from our analysis of the data considered in this work are highlighted, thus serving as a summary of the previous sections as well. In addition, here is shown how these findings can be interpreted, and if any patterns were identified.

7.1 Amount of errors

As mentioned earlier there is a relatively large number of errors introduced both by the ASR and the MT systems. A total of 1,860 errors was jointly identified in the ASR and MT outputs. Considering the total number of words transcribed (3688), this means that there is approximately one mistake in every two words. The MT system is the tool that introduces more errors (1220), thus having a 0.33 error per word rate. The ASR, with its 640 errors, performs at a 0.17 error per word rate. Due to the lack of reference material (for example error analysis in different languages) we cannot say whether or not this is a lot, but rates of 0.33 and 0.17 error per word are seemingly high error rates.

7.2 Independence between ASR and MT performance

In addition, the performance of the ASR system seems to have a major influence on the quality of MT outputs: it is very likely that an error in the ASR will also lead to an error in the MT output. This became apparent in earlier sections of this report where we show that of the 640 ASR errors, 566 led to an error in the MT output, which represents a percentage of 88.44%. At the same time, this means that of the 1220 MT errors 565 were caused by the ASR, which corresponds to a percentage of 46.31%.

Regarding the impact of specific types of ASR errors, *Punctuation* and *Named Entity* errors are among those with a closer relation to the performance of the MT system: 87.94% of all *Punctuation* errors lead to an error in MT; 90.48% of the ASR *Named Entity* errors also lead to a *Named Entity* error in MT. This probably has to do with the characteristics of this type of error. For example, if a name is wrongly transcribed by the ASR (*Named Entity*), it is very likely that the MT system will not be able to arrive at an accurate translation of what is actually said by the speaker. It makes sense that the MT has difficulties to find out if the name in the source text is the correct one and it is also statistically shown by the Le Mans research that a *Named Entity* error in the ASR often leads to *Named Entity* error in the MT (Gannay, et al., 2020).

This also applies to, for example, *Addition* errors and *Extraneous Word* errors. As the *Annotation Guidelines* show, an *Addition* error is described as an error in which text is added in the translation that is not present in the source text while *Extraneous* words are described as words that were not in the audio but were added by the ASR because the system was not able to transcribe it perfectly. So, if there is extraneous text in the transcription, which is the source text for the MT system, it is also likely that it will translate the extraneous text. This interdependence between these types of errors is made apparent by the amount of errors found in our data: all but one *Extraneous Word* errors lead to an *Addition* errors in MT outputs. Reversely, 9 of the 19 MT *Addition* errors were caused by an *Extraneous Word* error (47.47%). Thus, although almost all *Extraneous Words* lead to an *Addition* error, not all *Addition errors* are caused by an ASR *Extraneous Word* error. There is, nonetheless, also a close relation between MT *Addition* errors and ASR *Punctuation* errors, thus showing that most MT *Addition* errors are caused by the ASR.

These data make apparent that the ASR can have a major influence on the performance of the MT system. On the one hand, from the point of view of the ASR, an error in the ASR almost always causes an error in the MT output. From the point of view of the MT, almost 50% of the errors identified in MT outputs are caused by the ASR.

7.3 Most common errors in ASR and MT outputs

The analysis of the data shows that *Punctuation* errors, *L*exical *Selection* errors, *Grammar* errors, *Overly Literal* errors and *Incorrect Word* errors are the 5 most common types of errors overall. If we look at the ASR and MT tools separately, *Punctuation* errors and *Incorrect Word* errors are the most common types of error in ASR outputs and the *Lexical Selection* error, the *Punctuation* error, the *Grammar* error and the *Overly literal* error the most common in the MT system.

This allows us to realize that the *Punctuation* error is one of the most common errors in both the ASR and the MT outputs, which leads us to hypothesize that *Punctuation* errors are percolating from ASR outputs to MT's. To verify the validity of this assumption let us consider the percentage of ASR *Punctuation* errors that lead to a *Punctuation* error in MT: of the 423 *Punctuation* errors found in ASR outputs, 367 cause a *Punctuation* error in MT. This corresponds to 86.76% of ASR *Punctuation* errors and means that improving the performance of the ASR in terms of how it transcribes punctuation would have a major impact in the

performance of the MT system. This also shows the importance of an adequate use of interpunction in the source text to the performance of a MT system.

Focusing now on MT, *Lexical selection* errors are the most common with this tool. *Lexical selection* errors correspond to 34.10% of MT errors (416). Of these, 106 (25.48% of *Lexical selection* errors) were caused by the ASR. This is largely because *Incorrect word* errors are relatively common error in ASR outputs and these cause a *Lexical Selection* error in MT. This means that, besides having a significant impact on *Punctuation* errors in the MT, the performance of the ASR also has a significant influence on the quality of the translation itself produced by the MT system.

7.4 Severity of the errors

The data considered in the "Analysis and Results" chapter showed that *minor* and *critical* errors are the most common errors in terms of severity. Despite the fact that most of the errors were considered *minor*, a relatively large proportion of errors appear to cause significant problems: slightly more than 50% of the errors annotated in our data are considered *major* or *critical*. Respectively, the most common errors with a *minor*, *major,* and *critical* severity were *Punctuation*, *Lexical Selection,* and *Overly Literal* errors. Which also became apparent is that an error with a certain severity level does not necessarily have to be caused by an error with the same severity level: for instance, a *minor* error can easily cause a *critical* error. This occurred particularly in errors where specific features of Dutch also appear to play a role.

7.5 Role of the Dutch language

It also became apparent that some errors were due to specific features of Dutch.What should be emphasized again is the fact that despite the relatively low number of errors that can be attributed to asymmetries between English and Dutch, they seem to be types of errors that are likely to recurrently cause problems, as they are link to specific linguistic phenomena. This means that if we consider source texts in which these phenomena occur more often, there is a good chance that the number of such errors will also increase.

7.6 Relation between certain variables

During the analysis, we aimed at identifying connections between certain variables. Examples include relationships between *Lexical Density* or *Readability*, and the type of video,

the proficiency – or nativeness – of the speaker in English, and sentence length. We also checked whether there was a relationship between these variables and the number of errors found in the data. The analysis showed that certain patterns could be identified. For example, it seems that texts with longer sentences also tend to have a higher *Lexical Density* and be more difficult to read. However, no association was found between sentence length, *Lexical Density* and *Readability*, and the type of video and nativeness of the speaker in English. Even though most user content videos also had a low *Lexical Density* and a high *Readability* score, there were also user content videos that had a relatively high *Lexical Density*. Also, no evidence was found that the videos in which, for example, a non-native English speaker was present have a lower *Lexical Density* or higher *Readability* score.

When looking at these variables in relation to the type of errors or the number of errors that were made, a few patterns or tendencies could also be identified. So, it seems that user content videos are more prone to ASR errors. This is possibly not very surprising as we can consider that speakers in videos with a more informal context, such as user content videos, focus less on the pronunciation of words and sentences. As a result, it is possible that they speak faster, and that there is a greater chance of slips of the tongue or sentences that are suddenly cut, making it more difficult for the ASR to transcribe the text properly. For example, videos 7 and 8, in which speakers rate a product, have many ASR *critical* errors, such as *Incorrect Words*. This kind of video also tends to contain a lot of *crutch words* such as "ehm" and "okay". In video 8, for example, there are many MT *Should not be translated* errors (that is how *crutch word* errors are treated at MT level) in comparison with other videos. However, once again, full-proof evidence is lacking, because there are also user content videos where few errors of this nature occur. In addition, in terms of the number of errors identified, we realized that many videos with a high *Lexical Density* and low *Readability* score have a lower *Error to Word* ratio. However, the video with the lowest *Error to Word* ratio was a video with a low *Lexical Density* and high *Readability* score. This means that in this case the results do not allow us to categorically state whether there is a connection or not.

In general, even if the results are too variable to allow us to undoubtedly conclude that there are certain relationships between variables, they are, nonetheless, interesting for future work, as several possible connections seem to emerge from our analysis. The fact that some of the patterns or tendencies identified in our analysis showed significant variation in our data does not mean that there are no connections or patterns. That is why it is important, that subsequent studies are conducted with more data, and, where possible, in different languages, as described later in the "Future Work" section.

8. Conclusion

First, it should be said that, even though necessarily partial, this study gives a good picture of possible error patterns of ASR and MT systems, and of which types of error in which quantity are produced by both systems. Also, by studying the severity level of the errors annotated in our data, we were able to show which errors are the most problematic.

Even though the amount of data analyzed in this work is not enough to draw categoric conclusions, it allow us, nonetheless to outline trends of behavior. For example, these data show that the ASR appears to have a major influence on the performance of the MT system. If an ASR makes a mistake, there is a good chance that a mistake will also be made by the MT system. Since a significant proportion of ASR errors are *Punctuation* errors and cause problems in word order, many errors could be avoided by improving the ASR in this respect. Even if it is generally assumed that *Punctuation* errors do not cause many problems, the data analysis showed that ASR *Punctuation* errors lead to *critical* MT errors.

The data also show that there seem to be language-specific error patterns, in the case of our study, errors that can be related to specific properties of Dutch. Even though it involves only 25 instances, language-specific errors seem to be relatively regular and, thus, occur more frequently, depending on the characteristics of the source text. To address such errors, the *Annotation Guidelines* used at Unbabel could be extended and include language-specific instructions so that these kinds of errors could be detected and prevented.

Finally, our analysis also shows that there are possible relationships between certain text-specific variables, such as *Readability*, *Lexical density*, video type and so on. However, to be sure that these connections really exist, it is important to collect more data in multiple languages, if possible, which would provide more reliability to the picture sketched by our research.

9. Future work

As stated in the methodology section, this report presents exploratory research aiming at getting a picture of the errors made by ASR and MT systems used at Unbabel, in order to support future work in this area. To confirm the general trends and patterns identified in this work, multiple videos should be examined, in multiple languages, if possible. Extending the work presented here, focusing on the aspects that our analysis has shown to be relevant, will provide further evidence and serve to confirm or refute the patterns outlined by our research.

To get a more reliable image of the behavior of the data some changes should be introduced in the *Annotation Guidelines*, given that some errors were annotated as belonging to a certain type, although they could have been addressed more specifically. Also, our analysis as shown that, specifically for Dutch, there are language-specific error patterns. Accounting for this in the *Annotation Guidelines* would allow for getting more precise annotation data, rather than a general image. Doing so would possibly make way to a (semi-) automatic treatment of such errors. Also, if this has been observed for Dutch, it is likely that language-specific patterns can be identified for other language pairs, which opens a relevant line for future work.

10. Bibliography

ALPAC. (1966). *Language and Machines: Computers in Translation and Linguistics.* Washington D.C.: National Research Council.

Anastasiou, D. (2010, January 6). *Idiom Treatment Experiments in Machine Translation.* Retrieved from ResearchGate: https://www.researchgate.net/publication/294737858_Idiom_treatment_experiments_in_machine_translation

Arnold, D. (1994). Machine Translation: An Introductory Guide. London: NCC Blackwell Ltd.

Bahdanau, D., Cho, K., & Bengio, Y. (2016, May 19). *Cornell University.* Retrieved from Neural Machine Translation by Jointly Learning to Align and Translate: https://arxiv.org/abs/1409.0473

Bar-Hillel, Y. (1951). The present state of research on mechanical translation. *American Documentation*, pp. 229-237.

Bar-Hillel, Y. (1960). The present status of automatic translation. *Advances in computers*, pp. 91-163.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., . . . Monz, C. (2019, August 4). Findings of the 2019 Conference on Machine Translation (WMT19). *Association for Computational Linguistics*, pp. 1-61.

Benk, S., Dennai, A., & Elmir, Y. (2019). *A study on Automatic Speech Recognition.* Retrieved from ResearchGate: https://www.researchgate.net/publication/337155654_A_Study_on_Automatic_Speech_Recognition

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016, August 16). *EMNLP.* Retrieved from Neural versus Phrase-Based Machine Translation Quality: a Case Study: https://arxiv.org/abs/1608.04631

Besten, H. d., & Edmondson, J. A. (2002). *The verbal complex in continental west germanic.* Groningen: DBNL.

Bhat, D. N. (2007). Pronouns. Oxford: Oxford University Press.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., . . . Post, M. (2016). Findings of the 2016 Conference on Machine Translation (WMT16). *Association for Computational Linguistics*, 131-198.

Bourlard, H., & Morgan, N. (1994). Connectionist Speech Recognition: A Hybrid Approach. *Journal of Engineering and Computer Science*(247), 50-54.

Brinton, L. (2000). The structure of modern English. Amsterdam: John Benjamins.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J. D., . . . Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 79-85.

Bruderer, H. (1978). *Handbuch der maschinellen und maschineunterstützen Sprachübersetzung.* Munich: Vlg. Dokumentation.

Budiansky, S. (2000). Battle of Wits: The Complete Story of Codebreaking in World War II. New York: Viking Press.

Campbell, J., & Cuena, J. (1989). Machine translation, NLP, databases and computer-aided translation. *Perspectives in artificial intelligence*(2).

Carnegie Mellon University. (2005, November 18). *Microsoft.* Retrieved from Multi-Engine Machine Translation Guided by Explicit Word Matching: https://www.microsoft.com/en-us/research/video/multi-engine-machine-translation-guided-by-explicit-word-matching/

Castilho, S., Moorkens, J., Gaspari, F., Popovic, M., & Toral, A. (2019). Editors foreword to the special issue on human factors in neural machine translation. *Machine Translation*, 1-2.

Cohen, J. (2008, January 14). *The GALE project: A description and an update.* Retrieved from Institute of Electrical and Electronics Engineers: https://ieeexplore.ieee.org/document/4430115/authors#authors

Corbé, M. (1960, March 6). La machine à traduire française aura bientôt trente ans. *Automatisme*, pp. 25-30.

Corver, N., & Riemsdijk, H. v. (2001). Semi-lexical categories: The function of content words and the content of function words. Berlin: Mouton de Gruyter.

Costales, A. F. (2009). *The role of Computer Assisted Translation in the field of software localization.* University of Oviedo.

Crystal, D. (1980). Introduction to language pathology. London: Edward Arnold Ltd.

Cucu, H., Buzo, A., & Burileanu, C. (2014, September 3). *ASR errors in transcribing informal pronounciations of Romanian numbers.* Retrieved from https://pdfs.semanticscholar.org/fdf0/4a00f3ff87320c485065332fb0636b501b8a.pdf

Dahami, Y. S. (2012, January 25). *Adjectives and their Difficulties in English and Arabic: A Comparative Study.* Retrieved from ResearchGate: file:///Users/nickloomans/Downloads/AdjectivesandtheirDifficultiesinEnglishandArabicAComparativeStudy.pdf

Dale, E., & Chall, J. S. (1949). The Concept of Readibility. Chicago: National Council of Teachers of English.

Davis, K., Biddulph, R., & Balashek, S. (1952). Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America, 24*(6), 627-642.

Deng, L., Hassanein, K., & Elmasry, M. (1994, October 18). Analysis of the correlation structure for a neural predictive model with application to speech recognition. *Neural Networks, 2*(7), pp. 331-339.

Dictionary.com. (1995, May 14). *DOT.* Retrieved from https://www.dictionary.com/e/acronyms/dot/

Dingemanse, K. (2018, June 25). *Observatie in je scriptie - wanneer kan je het gebruiken?* Retrieved from Scribbr: https://www.scribbr.nl/onderzoeksmethoden/observatie-je-scriptie/

Dolan, W., Pinkham, J., & Richardson, S. (2002). MSR-MT: The Microsoft Research Machine Translation System. *Association for Machine Translation in the Americas*, 237-239.

Doll, J. (2012, September 12). A Literal Epidemic of Crutch Words. *The Atlantic*.

Dostert, L. (1955). The Georgetown-I.B.M. experiment. In W. Locke, & D. Booth, *Machine Translation of Languages.* Westport: Greenwood Press.

DuBay, W. (2006). Unlocking Language: The Classic Readibility Studies. Washington, D.C.: American Psychological Association.

Dudley, H., & Tarnozcy, T. (1950). The Speaking Machine of Wolfgang von Kempelen. New York: The Journal of the Acoustical Society of America.

Dudley, H., Riesz, R., & Watkins, S. (1939). A Synthetic Speaker. *Journal of the Franklin Institute*(227), 739-764.

Emmert-Streib, F. (2019). *Understanding statistical hypothesis testing: the logic of statistical interference.* Tampere: Multidisciplinary Digital Publishing Institute.

Farlex Inc. (2003, June 5). The Free Dictionary. Hundington Valley, Pennsylvania, United States.

Finch, J. (2000). Interprofessional education and teamworking: a view from the education providers. British Medical Journal.

Flesch, R. (1949). The Art of Readable Writing: With the Flesch Readibility Formula. New York: Harper & Row Publishers.
Formplus. (2007, December 7). *Formplus*. Retrieved from Exploratory Research: What are its Method & Examples?: https://www.formpl.us/blog/exploratory-research
Görög, A. (2017, June 23). *The 8 most used standards and metrics for Translation Quality Evaluation*. Retrieved from TAUS: https://blog.taus.net/the-8-most-used-standards-and-metrics-for-translation-quality-evaluation
Gürbuz, N. (2017). *Understanding fluency and disfluency in non-native speakers' conversational English.* Ankara: Middle East Technical University.
Gannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., & Morin, E. (2020). *End-to-end named entity and semantic concept extraction of speech.* Lyon: HAL (open archive).
Garg, A., & Agarwal, M. (2018, December 28). *Machine Translation: A Literature Review.* Retrieved from Cornell University: https://arxiv.org/abs/1901.01122
Godfrey, J., Hollman, E., & McDaniel, J. (2002, August 6). *SWITCHBOARD: telephone speech corpus for research and development.* Retrieved from Institute of Electrical and Electronics Engineers: https://ieeexplore.ieee.org/document/225858
Gompel, M. v. (2009, July 6). *Phrase-based Memory-based Machine Translation.* Retrieved from ResearchGate: https://www.researchgate.net/figure/The-Vauquois-triangle-Vauquois-1968_fig1_239924880
Google. (2005, May 22). *Google Translator: The Universal Language.* Retrieved from Outercourt.com: http://blogoscoped.com/archive/2005-05-22-n83.html
Gómez-Rodríguez, C., Caroll, J., & Weir, D. (2008). *Deductive approach to Dependency Parsing.* Brighton: Association For Computational Linguistics.
Haeseryn, W. (1984). Algemene Nederlandse Spraakkunst. Nijmegen: Martinus Nijhoff Uitgevers.
Halliday, M. (1985). Spoken and Written Language (Language Education). Oxford: Oxford University Press.
Harmer, J. (1997). How to teach English. Boston: Addison-Wesley.
HarperCollins. (2011). Collins English Dictionary. Glasgow, Glasgow, Schotland.
Haselow, A. (2017). Spontaneous spoken English: An integrated approach to the emergent grammar of speech. Cambridge: Cambridge University Press.
Hensel, T. (2014). *Validation of the Flesch-Kincaid Grade level within the Dutch educational system.* Twente : University of Twente.
Hoek & Sonépouse. (2012, March 22). *Richtlijnen voor Nederlandse ondertiteling.* Retrieved from Open Subtitles Forum: https://forum.opensubtitles.org/viewtopic.php?t=12639
Homer, D. (1940, October 6). The Carrier Nature of Speech. *Bell System Technical Journal XIX*, 508.
Huang, X., & Baker, J. (2014, January 18). *A Historical Perspective of Speech Recognition.* Retrieved from ResearchGate: https://www.researchgate.net/publication/262157198_A_Historical_Perspective_of_Speech_Recognition
Hulp bij Onderzoek. (2017, February 26). *Exploratief onderzoek*. Retrieved from Hulp bij Onderzoek: https://hulpbijonderzoek.nl/online-woordenboek/begrippen/exploratief-onderzoek/
Hutchins, J. (1978, June 2). *Machine Translation and Machine-Aided Translation.* Retrieved from Hutchinsweb: http://www.hutchinsweb.me.uk/JDoc-1978.pdf
Hutchins, J. (1992). An Introduction to Machine Translation. London: Academic Press.
Hutchins, J. (1999, September 13). Milestones in machine translation. *International Journal for Language and Documentation*, pp. 20-21.

Hutchins, J. (2001). *Machine translation over fifty years.* Retrieved on October 2020, from Hutchinsweb: http://hutchinsweb.me.uk/HEL-2001.pdf
Hutchins, J. (2005, January 18). *Towards a definition of example-based machine translation.* Retrieved from ResearchGate: https://www.researchgate.net/publication/228703010_Towards_a_definition_of_example-based_machine_translation
Hutchins, J. (2014). *The history of machine translation in a nutshell.* Retrieved on October 2020, from Hutchinsweb: http://www.hutchinsweb.me.uk/Nutshell-2005.pdf
Hutchins, J. (2016). Machine translation: a concise history. *Journal of Translation Studies*, 29-70.
IBM. (1961, September 6). *IBM Shoebox*. Retrieved from IBM: https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html
IBM. (2020, May 1). *Deep Learning*. Retrieved from IBM: https://www.ibm.com/cloud/learn/deep-learning
Irfan, M. (2017, January 28). *Machine Translation.* Retrieved from ResearchGate: https://www.researchgate.net/figure/Example-based-Machine-Translation_fig3_320730405
ISO. (2017, January 5). *Good Standardization Practices.* Retrieved from ISO: https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100440.pdf
Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. Piscataway: IEEE.
Jelinek, F., Bahl, L., & Mercer, R. (1975). Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. *IEEE Transactions on Information Theory*(21), 250-256.
Johansson, V. (2008). *Lexical diversity and lexical density in speech and writing: a developmental perspective.* Lund: Lund University .
Joshi, M. (2013). English interrogative sentences: common interrogative patterns. Scotts Valley: CreateSpace Independent Publishing Platform .
Juang, B., & Rabiner, L. (2005, January 8). *Automatic Speech Recognition - A Brief History.* Retrieved from ResearchGate: https://www.researchgate.net/publication/249888949_Automatic_Speech_Recognition_-_A_Brief_History_of_the_Technology_Development
Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the Association for Computational Linguistics*, 1700-1709.
Kang, J. (2019, March 6). *What is User Generated Content and How It Is Relevant?* Retrieved from Business2Community: https://www.business2community.com/content-marketing/what-is-user-generated-content-and-how-it-is-relevant-02175516
Katz, S. (2019). American English Grammar: An Introduction. Abingdon: Routledge.
Kenny, D. (2018). Machine Translation. *The Routledge Handbook of translation and philosophy*, 428-445.
Kincaid, J. P. (1975). Derivation of New Readability Formulas: (automated Readibility Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Memphis: Naval Air Station Memphis.
Klare, G. R. (1963). The Measurement of Readibility . Ames: Iowa State University Press.
Klatt, D. H. (1977). Review of the DARPA Speech Understanding Project. *The Journal of the Acoustical Society of America*(62), 1345-1366.
Koamé, J. B. (2010, August 8). *Using Readibility Tests to Improve the Accuracy of Evaluation Documents Intended for Low-Literate Participants.* Retrieved from Journal of MultiDisciplinary Evaluation: journals.sfu.ca

Koehn, P. (2016, November 30). *Omniscien Technologies.* Retrieved on October 2020, from The State of Neural Machine Translation: https://omniscien.com/blog/state-neural-machine-translation-nmt/

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation . *First Workshop on Neural Machine Translation 2017*, 28-39.

Koehn, P., Birch, A., & Callison-Burch, C. (2007, June 19). *ResearchGate.* Retrieved on October 2020, from Moses: Open Source Toolkit for Statistical Machine Translation: https://www.researchgate.net/publication/220874004_Moses_Open_Source_Toolkit_for_Statistical_Machine_Translation

Koster, J. (2002). *Dutch as an SOV language* . Den Haag: DBNL.

Kratzenstein, C. G. (1782). Sur la naissance de la formation des voyelles. Saint Petersburg: Saint Petersburg Academy of Sciences.

Lai, J., Karat, C., & Yankelovich, N. (2008). Converstational speech interfaces and technologies. In J. Jacko, & A. Sears, *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (Vol. 2nd edition, pp. 381-91). New York: NY: Erlbaum.

Lane, D. M. (2013). Introduction to Statistics: An Interactive e-Book. Houston: Rice University.

Lee, S., & In, J. (2015). *Standard deviation and standard error of the mean.* Seoul: Korean Journal of Anesthesiology.

Lemmens, M., & Slobin, D. I. (2007). *Position verbs and movement verbs in the Dutch, the English and the French language.* Lille: Lille University.

Liu, R. (2018, March 4). *Translation Quality Assessment: MQM (Multidimensional Quality Metrics)*. Retrieved from Middlebury Institute site network: https://sites.miis.edu/runyul/2018/03/04/translation-quality-assessment-mqm-multidimensional-quality-metrics/

Localization Industry Standards Association. (2006, September 9). *LISA QA Model 3.1 -- Assisting the localization development, production and quality control processes for global product distribution.* Retrieved from DSS Resources: http://dssresources.com/news/1558.php

Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., & Popovic, M. (2015, July 31). *DQF & QT21.* Retrieved from TAUS: http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf

Markoff, J. (2012, November 30). Scientists See Promise in Deep-Learning Programs. *New York Times*, pp. 25-26.

Martins, A. (2019). *OpenKiwi: An Open Source Framework for Quality Estimation* . Retrieved from Unbabel: https://medium.com/unbabel/openkiwi-an-open-source-framework-for- quality-estimation-30c35a998a9f

Martins, A., Almeida, M., & Smith, N. A. (2013). *Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers.* Sofia: Association for Computational Linguistics.

Maučec, S., & Donaj, G. (2019, September 7). *Intechopen.* Retrieved on October 2020, from Machine Translation and the Evaluation of Its Quality: https://www.intechopen.com/books/recent-trends-in-computational-intelligence/machine-translation-and-the-evaluation-of-its-quality

McArthur, T. (2011). The Oxford companion to the English Language. Oxford: Oxford University Press.

McCaskill, M. K. (1998, August 3). *Grammar, Punctuation and Capitalization: A Handbook for Technical Writers and Editors.* Retrieved on December 2020, from Rose-Hulman Institute of Technology: https://www.rose-hulman.edu/class/ee/HTML/ECE340/PDFs/grammar_NASA.pdf

McLean, J. E. (1998). Statistical Significance Testing. Alabama: University of Alabama.
McWhorter, J. (2016, November 25). The Evolution of 'Like'. *The Atlantic*.
MDPI. (2019, March 18). *The Rapid and Steady Mass Loss of the Patagonian Icefields.* Retrieved from Multidisciplinary Digital Publishing Institute: https://www.mdpi.com/2072-4292/11/8/909/pdf-vor
Melčuk, I., & Ravič, R. (1967). Automatic translation: 1949-1963: A critical bibilographic reference. Moscow: All-Union Institute for Scientific and Technical Information.
Merriam-Webster Incorporated. (1999). Merrian-Webster's Collegiate Dictionary. Springfield, Massachusetts, United States .
Mohri, M. (1997). Finite-State Transducers in Language and Speech Processing. *Computational Linguistics, II*(23), 269-312.
Morris, C. (1992). Academic Press Dictionary of Science and Technology. Houston: Gulf Professional Publishing.
Moskvitch, K. (2017, February 15). *The machines that learned to listen*. Retrieved from BBC: https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen
Nafayan, M., Safayi, S., Hansen, J., & Russel, M. (2016). *Improving Speech Recognition using limited accent diverse British English training data with deep neural networks.* Salerno, Italy: University of Salerno.
Nagao, M. (1984). A framework of a Mechanical Translation between Japanese and English by Analogy Principle. *Artificial and Human Intelligence, A. Elithorn and R. Banerji, eds.)*, 173-180.
Neijt, A., & Hoekstra, H. (1986). Vertalen per computer. *De Gids. Jaargang 149*.
Nesia, B. H., & Ginting, S. A. (2014). *Lexical density of english reading texts.* Medan: University of Medan.
Nida, E. (1984). On Translation. Beijing: Translation Publishing Corporation.
O'Brien, S. (2012, October 21). *Towards a Dynamic Quality Evaluation Model for Translation.* Retrieved from The Journal of Specialized Translation : http://www.jostrans.org/issue17/art_obrien.php
Oliveira, T. R. (2019). *Análise da tradução de combinatórias lexicais em tarefas de pós-edição de tradução automática.* Lisbon: Universidade de Lisboa.
Oxford Languages. (2009). Oxford English Dictionary. Oxford, Oxfordshire, England.
Parra, G. (2005, February 28). *La Revisión de Traducciones en la Traductología: Aproximación a la Práctica de la Revisión en el Ámbito Profesional Mediante el Estudio de Casos y Propuestas de Investigación.* Retrieved from University of Granada: http:// digibug.ugr.es/handle/ 10481/660
Peng, H. (2018). *The impact of Machine Translation and Computer-aided Translation on Translators.* Wuhan: Wuhan Polytechnic University.
Pijarnsarid, S. (2017). *An Analysis of the Content Words Used in a School Textbook.* Ubon Ratchathani: Ubon Ratchathani Rajabhat University .
Pinola, M. (2011, November 2). *Speech Recognition Through the Decades: How We Ended Up With Siri* . Retrieved from PcWorld: https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html
Poai, S. (2012). *Students' mastery in using adverbs at english study program of sintuwu maroso university.* Poso: Sintuwu Maroso University.
QT21. (2012, May 5). *The New Goal of Quality Translation*. Retrieved from QT21: https://www.qt21.eu/launchpad/
QT21. (2015, December 30). *Multidimensional Quality Metrics (MQM) Issue Types.* Retrieved from QT21: https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html

Rabiner, L. R. (1989, February 16). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE 77, 2*, pp. 257-286.
Rensink, R. A. (2017). *The nature of correlation perception in scatterplots.* Vancouver.
Rojo, J. L. (2018). *Aspects of human translation: The current situation and an emerging trend.* Malaga: University of Malaga.
Samudravijaya, K. (2008, September 8). *Indian Institute of Technology Guwahati.* Retrieved from Automatic Spech Recognition: http://www.iitg.ac.in/samudravijaya/tutorials/asrTutorial.pdf
Schwartz, L. (2016). *History and Promise of Machine Translation.* Illinois: University of Illinois.
Schwartz, L. (2016). *The history and promise of machine translation.* Illinois: University of Illinois.
Shetter, W., & Ham, E. (2007). Dutch: An Essential Grammar. London: Routledge.
Shrestha, A. (2019, April 1). *Review of Deep Learning Algorithms and Architectures.* Retrieved from Research Gate: https://www.researchgate.net/figure/GNMT-Architecture-84-with-encoder-neural-network-on-the-left-and-decoder-neural-network_fig16_332662087
Somers, H. (1999). Review Article: Example-based Machine Translation. *Machine translation*, 113-157.
Stankevičiūtė, G., & Kasperaviciene, R. (2017). *Issues in Machine Translation.* Kaunas: Kaunas University of Technology.
Stebbins.R, & Publications, S. (2001). *Exploratory Research in the Social Sciences.* Thousand Oaks, Canada: SAGE Publciations.
Subrahmanyam, J. (2012). Current English Grammar and Usage. Chennai: Sura Books Pvt Ltd.
Svendsen, T. (2003). Speech Technology: Past, Present and Future. Oslo: Telektronikk.
Swaen, B. (2019, January 17). *Wat is kwalitatief en kwantitatief onderzoek?* Retrieved from Scribbr: https://www.scribbr.nl/onderzoeksmethoden/kwalitatief-vs-kwantitatief-onderzoek/
Swamy, S., & Ramakrishnan, K. (2013, July 6). *Evolution of Speech Recognition – A Brief History of Technology Development.* Retrieved from Elixir Publishers: https://www.elixirpublishers.com/articles/1373104268_60%20(2013)%2016239-16243.pdf
TAUS. (2015, June 15). *DQF and MQM harmonized to create an industry-wide quality standard*. Retrieved from TAUS: https://www.taus.net/academy/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard
Taysk, R. (2019). *Qualidade na tradução automática e na pós-edição: anotação de erros de concordância e ordem de palavras.* Lisbon: Universidade de Lisboa.
Testa, I. (2018). *Quality in human post-editing of machine translated texts: error annotation and linguistic specifications for tackling register errors.* Lisbon: Universidade de Lisboa.
To, V., Fan, S., & Thomas, D. (2013). *Lexical Density and Readibility: A Case Study of English Textbooks.* Tasmania: University of Tasmania.
Trojanskij, P. P. (2000). Description of a machine for selecting and typing words when translating from one language into another or several simultaneously. *Machine Translation*.
Unbabel. (2019, September 29). *Distributed pipeline.* Retrieved from Unbabel: https://unbabel.com/
Unbabel. (2019, September 29). *Distributed pipeline for video.* Retrieved from Unbabel: https://unbabel.com/

University of Cambridge. (2002, April 15). *HTK Rich Audio Transcription Project Summary and Aims*. Retrieved from Department of Engineering of University of Cambridge: http://mi.eng.cam.ac.uk/research/projects/EARS/ears_summary.html
Ure, J. (1971). Lexical density and register differentiation. London: Cambridge University Press.
Utah Valley University. (1999). *Types of Verbs.* Retrieved from Utah University: https://www.uvu.edu/writingcenter/docs/handouts/grammar/typesofverbs.pdf
van Bregt, A. (2012, September 24). *Verschillen in marketing voor persoonlijke of professionele netwerken*. Retrieved from Social Media Academie: https://www.socialmediaacademie.nl/verschillen-in-marketing-voor-persoonlijke-of-professionele-online-sociale-netwerken/
Van Dale Uitgevers. (2015). Van Dale Groot woordenboek van de Nederlandse taal. Utrecht, Utrecht, The Netherlands.
Waibel, A., & Lee, K.-F. (1990). Readings in Speech Recognition. Amsterdam: Elsevier Science & Technology.
Warriner, J. (1981). Warriner's English Grammar & Composition. Harcourt: Harcourt School Publishers.
Way, A. (2018). *Quality expectations of machine translation.* Retrieved from https://arxiv.org/pdf/1803.08409.pdf
Weaver, W. (1949). Translation. In W. Locke, & A. Booth, *Machine translation of languages: fourteen essays* (pp. 15-23). Massachusetts: Massachusetts Instititue of Technology.
Williams, T. (1940, January 25). At the New Yorks World's Fair. *Bell Telephone Quarterly, 19*(1), pp. 59-71. Retrieved from Internet Archive.
Winkler, E. G. (2008). Understanding Language: A Basic Course in Linguistics. New York: Continuum.
Wolk, K., & Marasek, K. (2015). Neural-based Machine Translation for Medical Text Domain. *Procedia Computer Science*.
Yule, G. (1999). Explaining english grammar. George Yule: Oxford University Press.