University of Copenhagen
Faculty of Humanities
Department of Nordic Studies and Linguistics

Name of department: Department of Nordic Studies and Linguistics

Name of programme: IT & Cognition

Author(s): Jakob Blaaholm Nielsen, tjl698

Master's Thesis

Evaluation of Machine Translations from Google Translate, eTranslation and DeepL: A Quality Assessment of the Machine Translations from English to Danish and vice versa

Supervisor: Bolette Sandford Pedersen

Submitted on: 3rd of April 2022

No. of characters:130569

No. of pages: 53

Acknowledgments

I would like to thank my supervisor Bolette Sandford Pedersen for her advice, feedback and for helping establish contact to Claus Thornby Larsen and the Agency of Digitalisation, where I was able to sit at an office space for most of the writing period. In this context, I would like to thank Christian Plaschke, Josephine Worm Andersson and Carl Frederik Bach Kirchmeier for their sparring and general helpfulness and for letting me ask the obvious questions and them answering the questions. Also, a big thank you to Claus Thornby Larsen who has enlightened me on the curiosities of machine translations in the EU. I would also like to extend my gratitude to Mathias Ennegaard Asmussen for his helpfulness and patience, not to be forgotten. Also, Natalie Sørensen who helped read and provide feedback on various parts of the thesis. Finally, I would like to thank my family and friends for their lovely support during the production of my thesis.

Abstract

The purpose of this study is to benchmark Danish to English machine translations and vice versa and to estimate if eTranslation could be a viable alternative to Google Translate or DeepL. Stakeholders within Danish NLP have identified automatic translation as an area that can be improved and yield great value in return. I have therefore evaluated the quality of machine translations from eTranslation, Google Translate and DeepL using two evaluation metrics namely the BLEU and TER score and a custom error typology. This was done in an attempt to disprove the narrative that Danish machine translation is low quality and has not seen the same rise in quality as other languages since the introduction of neural networks in 2015. The comparison was made on four different domains to ensure a broad quality estimation. My findings indicate that Danish machine translations produced are high-quality. The three tools are quite similar in quality across domains and languages, but differ in other areas such as certain error types and post-editing effort. Given the similar high quality of the three tools, a user's decision can then be made on factors around the translation system, such as values and data security.

Content

1 Introduction

1.1 Motivation

In the report ’Sprogteknologi i verdensklasse’(Kirchmeier et al. 2019), the status of Danish language technology was given and a bleak picture was painted. One of many identified areas with potential for improvement was machine translation. It was mentioned that although the quality has soared since the introduction of neural networks, there is still a gap between the standards of human translations and machine translations at the time of completing the report. There has also been an increasing focus on handling personal information safely and responsibly, which makes the European translation tool, eTranslation, particularly interesting to evaluate. In addition, translation tasks are still in demand by public and private stakeholders alike, but these tasks have been increasingly outsourced to either professionals or the most readily available tool, namely Google Translate or DeepL (Kirchmeier et al. 2019, p. 28).

Therefore, I would like to study the quality of Danish to English machine translation today. Secondly, I examine whether eTranslation could be a realistic alternative to Google Translate or DeepL regarding Danish translation tasks. Also, is there a significant difference in the quality of translations and would a possible discrepancy be dependent on the domain? Concerning this: What does the choice of evaluation metrics mean for the total quality assessment?

My motivation for this is the increasing focus and funding into Danish language technology and solutions aimed at the Danish language in recent years. I want to contribute to the development of this Danish language technology. At the time of writing, private and public stakeholders have recognised some of the issues raised in the report and language technology is picking up speed in Denmark. There are initiatives such as the public website ‘sprogteknologi.dk’ launched by the Agency of Digitalisation that aim to gather and collect existing and future language resources, models, tools, corpora etc. to create a foundation for all stakeholders with an interest in Danish language technology. This was to face one of the previous challenges, namely that the effort to improve the Danish language technology has been disjointed, resulting in an idle state of development. For example, a student in need of a named entity recogniser or a different language resource would not have to reinvent the wheel, but simply click onto the sprogteknologi.dk webpage and find the most suitable solution for his or her needs.

Another challenge Danish language technology faces is the lack of high-quality linguistic data appropriate for tasks. To put Danish language technology into service, corpora with a

great amount of metadata holding information on for example pronunciation for speech systems, domains specific words or phrases for classification tasks and so on. A common denominator for information-enriched corpora is that the higher the quality demands are the higher the need for a linguistic expert to be involved is. These are not only expensive but also in short supply in Denmark. As it stands right now, the only master's programme that truly focuses on language technology is IT & Cognition at the University of Copenhagen, even though there are electives and other minor subjects to be taken at other institutions. There is another issue in relation to this, namely that the students do not work with Danish language data per se, but have to opt to examine Danish data at exams and tests etc. The scarcities of high-quality data and specialists might stem from the lack of a focal point and incoherent efforts to create a solid foundation like a national term bank. This is pointed out in the report as one of the top priorities and establishing a national language bank would not only save funds but also add to the limited amount of useful data available at the moment (Kirchmeier et al. 2019).

An initiative like this could also be the catalyst to illustrate the importance and value of the aforementioned data. At the time of writing, there are no standard procedures or systematic approaches for how municipalities choose to solve their translation tasks. The current practice is a mix of employees who know the target language but are employed for different tasks, others use external translators and only a very small portion use employees who are actively working with language-related tasks. The absence of structure prognosticates that no thoughts have been given to how the translated data is stored or what uses a translated text might have to others (Kirchmeier et al. 2019). There is no feedback of data to any storage or filing system and many valuable datasets are lost. To enhance the Danish field of machine translation, these texts could be stored in a translation memory. A translation memory is vital for training and is used by nearly all translation companies and is typically combined with machine translation. This could add a good chunk of data to the pool. Another option, which is readily available to public institutions and small and medium-sized enterprises, is the translation tool from the EU mentioned earlier, namely eTranslation. The conclusion of a workshop consisting of participants from public and private institutions, unions and representatives from the ELRC was that the trend of outsourcing public translation tasks could contribute to the state of language technology in Denmark if the data and not only the output were to be shared. However, a public-private collaboration would require regulation of data treatment and a re-examination of public tendering rules (Kirchmeier et al. 2019).

This point is made more relevant from a conclusion in the report shows that translation is a growing need for both public and private entities that are active across the borders of EU countries and since the EU aspire to encourage digital inner market trading, a tool that allows free translation into 24 languages is of high value. The European Language Resource Coordination (ELRC) is a collection of language resources made available for all member states of the EU with Norway and Iceland included as well (European Language Resource Coordination 2020). The Danish Language Council and the two Danish anchor points in the ELRC, The Agency of Digitalisation and the Centre for Language Technology at the University of Copenhagen have recognised this. They have contributed to the initiative ELRC by gathering parallel corpora from public institutions and making them available for training models. These models could be eTranslation's own or perhaps Danish-developed ones in the future. This effort has resulted in a steady increase in eTranslation's quality and improvement on smaller languages like Danish and could be seen as a possible gateway to more data if the tool could translate texts to a satisfactory level (European Language Resource Coordination 2020, p. 33). Furthermore, there has been an increasing amount of attention to how personal data is treated online in recent years, with executive vice-president and EU antitrust chief Margrethe Vestager from her position as commissioner for Competition in the European Union raising the alarm for a more secure treatment of personal data and point fingers at Google and Facebook (Breinstrup 2016). One of the cornerstones of eTranslation is the treatment and protection of personal information and data security. Although this does not set eTranslation in a position to be directly compared to the two most used and readily available tools Google Translate and DeepL, it does announce itself as possible viable alternative, exactly because it offers a new dimension, namely protection.

A comparison between eTranslation, Google Translate and DeepL might inspire a new preference for translation tasks in the public and private sectors depending on the outcome. In addition, it might reveal that Danish language technology is not that far behind if not a forerunner as Danish is in other parts of digital governance.

To perform the comparison and assess the quality of each translation tool, I am using a selection of corpora from the ELRC across different domains to evaluate eTranslation, Google Translate DeepL. The corpora from ELRC are converted from XML-files into .txt-files to enable the analysis and comparison of the sentences. The domains are Public Health, Culture, Finance and general text. General text is different due to its lack of a constraining vocabulary, thus offering a challenge to eTranslation, Google Translate and DeepL.

EU's eTranslation is a result of former machine translation services created by the European Commission MT@EC. This was in turn built upon an open-source toolkit for translation called MOSES. Whereas MT@EC was a statistical machine translation, eTranslation is a product of its time and is a neural network. In addition, the aim of eTranslation is scalability and flexibility to accommodate the potential growth in use (Connecting Europe Facility). Google Translate is also a neural network, making the transition in 2016 from a statistical, phrase-based machine translation model. Given the huge amount and influx of data that Google receives, its models are well fed and can present very accurate translations in many languages (Aiken 2019). DeepL is built upon an online dictionary called Linguee that scraped texts but is now a neural network claiming to be four times better than Google Translate. Because DeepL has a limit of 5000 characters, a number of sentences will be translated and evaluated; raising the character limit is an option to buy. This will be further elaborated on later.

The most used metric for evaluating machine translations for the past many years has been the Bilingual Evaluation understudy-score also called the BLEU score. This method is used to evaluate eTranslation and Google Translate uses its own GLEU-score. In my thesis, I will employ the BLEU score for both systems and I will also be using translation edit rate (TER) to estimate the quality of translated documents to avoid limiting myself to only one measurement. BLEU is a comparison of sentences that returns a score that indicates how identical the hypothesis translation is to the reference translation(s). TER is a registration of modifications/edits required for a candidate translation to mirror a reference sentence.

2 Background in NLP and Machine Translation

In this section, I will give some background knowledge on the status of Danish language technology and machine translation.

2.1 NLP in Denmark and beyond

Natural language processing (NLP), natural language generation, natural language understanding, text-to-speech, chat bots, conversational AI, and machine translations are just some of the terms included under the ever-growing umbrella of language technology and computational linguistics. Advances in processing power and data availability have resulted in a plethora of opportunities with these technologies. As mentioned beforehand, language technology has come a long way. Even though the print press originated in China centuries before Johannes Gutenberg's print presser saw the light of day around the middle of the 15th century, it can still be seen as one of many

instalments in a long line of technology initiatives that have helped produce, generate, understand and learn language and meaning. Fast-forward a couple of centuries and the famous Alan Turing, who gave name to the Turing Test, which sets out to determine whether a machine is intelligent or at least can simulate intelligence. I am not going to discuss the validity of the test, but I want to note that the interesting part is the question asked and whether such a thing as artificial intelligence exists. To take this idea further, you can ask if a machine is able to think instead of just displaying intelligence.

In almost all literature and movies aimed at either the future and/or space adventures, there has been some sort of way of verbally communication from humans to a central brain or a nexus of sorts, which displays various degrees of passive compliance or active defiance depending on the author's intended message. The computer has had the role of the villain or just as a household technology that would assist in everyday family life. As you read this, you might remember that you are low on milk and oats and utter the words 'Hey Siri – Can you add milk and oats to my shopping list?', which is confirmed either with a beep or a voice repeating your command. The future is now if you are from a well-sourced language domain. Native English speaking countries can reap the rewards of what the modern world has to offer, but other countries have to wait years to even see this technology and the possibilities that follow. Danish is one of those low-resource languages that have to wait, adapt or develop the technology themselves.

Language technologies have the potential to assist humans in many situations, like voice-controlled user interfaces, educational purposes and so on (Pedersen, Rehm, and Uszkoreit 2012). Currently, inclusion and accessibility are prioritised by companies and public sector entities in Denmark, which entails government websites to be made available to people with reading disabilities or otherwise struggle to access information. At Gyldendal Uddannelse, the educational department at the publishing company Gyldendal, all learning materials, images and videos have to be created in a format that can be read aloud by the computer or an application. Usually, these reading applications are imported, like the reader from Amazon or Mozilla. They work to a somewhat satisfactory level but do not fully understand Danish, which is a problem.

In order to develop Danish solutions, we would need a stronger foundation as mentioned in the introduction. There would be both commercial and intangible gains when developing better Danish models. Oddly, we are so far behind on this sort of technology considering Denmark is one of the most digital countries in the world. According to the Digital

Economy and Society Index (DESI) 2021 (Digital Economy and Society Index 2021), the digital infrastructure in Denmark ranks top among the EU countries. This entails availability of internet access, 5G readiness and so on. Furthermore, the report puts Danish small and medium enterprises in 1st place when ranking SMEs with a basic level of digital intensity, meaning they employ at least four digital technologies that 'enable businesses to gain competitive advantage, improve their services and products and expand their markets.' These could be big data, cloud solutions or the loosely defined AI (Digital Economy and Society Index (DESI) 2021). The Danish market is ready for high-quality solutions and the Danish digital government has also been promoted as one of the best globally.

In Denmark, digital governance has been a priority since the 1990s when digitalisation, digital strategies and several other initiatives were launched. In the following years, Denmark has benefited in a number of ways from these and has ranked at the top of the list regarding public sector digitalisation. Public health, taxes and communication between citizens and public authorities have all been boosted by these initiatives. In April 2019, stakeholders from the Danish government, the Danish Language Council and other peers published 'Sprogteknologi i verdensklasse'. The report mapped the current state of Danish language technology, what challenges lie ahead and gave suggestions to which areas within the field of language technology should be prioritised. These suggestions were based on surveys and workshops with researchers, developers, end-users and suppliers.

An initiative to aid the development of Danish language technology is sprogteknologi.dk[1]. It is a website created in 2020 by the Centre of Data and Technology a subdivision of the Danish Agency of Digitalisation. Metadata on Danish language resources and tools are available to use for everybody with an interest in Danish language technology and artificial intelligence. These are named entity recognisers and language models like BERT and Danish Electra, which are some of the state-of-the-art models in Danish. In their own words, 'the primary goal is to support the development of artificial intelligence in Danish and to make sure that the digital language in Denmark is Danish'. The metadata and resources are continuously collected in an agile manner to conform to the user's needs, which is a reflection of the rapid development happening in the field. There is also a political motive behind the initiative agreed upon by the government, KL – Local Government Denmark and Danish Regions. In addition to the employees

[1] https://sprogteknologi.dk/

at the Centre of Data and Technology, a panel of experts from various organisations, like the University of Copenhagen, the Danish Language Council and Rigshospitalet, just to name a few, have also been included in the steering group to give advice.

The steering group and other stakeholders call for better practices for gathering data and lobbying for a legal framework that makes data sharing easier for Danish NLP users. This would benefit the development of a foundation to further boost Danish NLP created by Danes with Danish principles. The European Language Coordination have their repository with parallel corpora amongst other readily available resources. This has been used to train their tool eTranslation. Machine translations have progressed to a point beyond the unintelligible, often laughable, attempts at translating snippets of text in one language into another, either for students or for professionals to save time on assignments. The quality of machine translations surpassed an adequacy threshold at some point, which has actually made them useful and they are now an essential tool for translators and other stakeholders in different domains like law or public health or even just general speech and text.

2.2 Related work – State of the art

Machine translations have been around for a long time. The automatic machine translation systems we see now are the product of many years of research and attempts at breaking the language barrier. Historically, there have been some major shifts that need to be mentioned. The first shift was from rule-based machine translations to example-based around 1980. Rule-based systems were the initial translation systems and encapsulate a pragmatic approach to translation. They saw the light of day around 1950. They are constructed by an expert who makes a set of linguistic and grammatical rules and structures and the machine is taught a vocabulary in both languages, which is very time-consuming. This approach works well for everything unambiguous, but not many sentences contain words without several interpretations and thus the output is generally of a low standard (Sepesy Maučec and Donaj 2020). Example-based systems also signify a straightforward approach to translation. Example-based systems are corpus-based, meaning they have bilingual corpora at their disposal. As the name suggests, the system dives into a corpus to locate examples that match the given input and finds the corresponding sentence that matches. The output is thus a patched version of sentences that match in some way or another.

In 1990, statistical machine translation systems gained ground and are still widely used today. They also incorporate corpora in their approach to translation and the training phase.

They are trained on a large corpus consisting of gold standard translations and from this are able to create a statistical translation model. This is essentially a table of phrase frequency, which logs how often a phrase is encountered throughout the entire corpus. The probability of that phrase being correctly translated in the first place increases with the number of times it is put into the table. The probabilistic design works very well when a lot of data is available. For an overview of the performance of statistical machine translation, see figure 3 in the appendix. This is both a strength and a weakness since it is scalable but also vulnerable to data sparsity. Large amounts of training data are also a challenge faced by neural machine translations. The newest shift happened around 2015 as major machine translation services started transitioning to this kind of translation. Neural machine translations is using neural networks to train a statistical model for automatic translation.

In theory, neural machine translations simulate how a brain works and learns to analyse text as it is presented with more data. In a sense, it is trained to recognise relationships amongst a large amount of data through various algorithms. A simple neural network consists of an input layer, a hidden layer and a target layer. These layers are connected by nodes to simulate the brain-like structure. When given an input, neurons calculate where the strongest relationship is and gives the information to the next neuron. Typically, the neurons have a certain value that determines whether the information should be passed or not. The input is looked at not sequentially but as a whole. The increasing amount of hidden layers in the simple neural network starts to form a deep neural network. In essence, a deep neural network teaches a computer how to solve a problem instead of telling it how to solve it. This can be used to train computers for a number of tasks that resemble intelligence or at least a higher level of cognitive ability than the basic statistical model. One of these tasks could be image processing or speech recognition or even machine translation. The way this training happens is also a vulnerability as spamming incorrect answers might disturb the algorithm and might be accepted as the correct answer and thus creating a new reality. For instance, when ‘Donald Trump’ appeared on Google when you searched for ‘idiot’. There is a black box element to the hidden layers as they are difficult to correct once the damage has been done (Wu et al. 2016).

In order to elevate the quality of natural language processing, the concept of ‘Transformers’ and ‘attention’ were introduced. This was done to solve some of the problems with ambiguity, where the confusing word is identified along with the most important other elements that help clear up the ambiguousness. In a Danish context, some of the most recent influential

transformer models are Danish BERT[2], Ælæctra[3] and Danish Røberta[4]. The Ælæctra-model is trained on the Danish GigaWord Project[5] and requires much fewer resources for training and processing. This shows that there are possibilities for Danish NLP (Vaswani et al. 2017).

The transition to neural machine translation has shown an improvement in the quality of machine translations. In a study on post-editing, Koponen et al. report that a comparison between neural machine translation and statistical machine translation showed 'an overall reduction of errors as well as a reduction in specifically morphological errors and word order errors in various language pairs' (Koponen, Salmi, and Nikulin 2019). Furthermore, in an update on the evaluation of Google Translate translations and which languages perform best and worst, the top 10 were: German, Afrikaans, Portuguese, Spanish, Danish, Greek, Polish, Hungarian, Finnish, and Chinese. This study was done using a variety of evaluation metrics. Thus, it might be as bleak a picture as has been painted for Danish machine translation (Aiken 2019).

2.3 Translation tools

In this section, I explain why I have chosen the three translation tools I have for the analysis.
I chose eTranslation, Google Translate and DeepL as the translation tools for this analysis for different reasons. First, the three tools are all free to use, although DeepL does not allow document translation for enough documents to do this examination without paying a subscription. As mentioned previously, eTranslation was highlighted as a tool for public sector translation due to data security and a reported rise in quality. I chose Google Translate as one of the tools because it is so widely used and has always been the tool to beat regarding the quality of free online translation. According to the recently published European Language Industry Survey 2022 (ELIS Research 2022), DeepL is the preferred choice by language companies. Furthermore, DeepL claims to be much better than market competitors such as Google Translate. In addition, DeepL offers translations in a good number of languages.

2.3.1 eTranslation

Built upon a framework of the predecessor, MT@EC, eTranslation is the product of the natural progression towards neural machine translations, which is the new black within the field. The Connecting Europe Facility's (CEF) eTranslation has been trained on various domains, comprising

[2] https://sprogteknologi.dk/dataset/ebdcd8fc-49ff-406a-83d8-2232aad95d0d
[3] https://github.com/MalteHB/-l-ctra
[4] https://huggingface.co/flax-community/roberta-base-danish
[5] https://gigaword.dk/

1 billion sentences from the Euramis translation memory. eTranslation offers translation between all the languages of countries part of the European Union, English, Norwegian and now also features Arabic and simplified Chinese resulting in +30 languages. The corpora's origin are EU documents translated by translators connected to the EU institutions and thus the domains offered by eTranslation all relate to the nature of official EU articles ('What Is ETranslation'). One of the strengths of eTranslation is exactly that of domain-specific translations, which, in theory, should offer more accurate translations in specialised areas such as law and public formal language. It is worth noting that in almost all translations not involving English, English is still used as a pivot language, meaning that a Danish to Polish translation would be translated from Danish into English and after that from English to Polish. This is a normal practice when data between language pairs is limited, as there usually exist parallel corpora between English and Danish and English and Polish. This feeds a different discussion on the growing amount of Anglicisms in otherwise 'pure/clean' translations and the quality hereof when introducing a third language to bilingual translation (Benjamin 2019). eTranslation is available, but not available to everybody; however many stakeholders can get access. Freelance translators for the EU, students, public administrations, territorial management units and recently small and medium-sized enterprises have access to the tool as long as they create a log in.

2.3.2 Google Translate

Although Google Translate does not have any domains, it supports over 100 languages and still adds to that list. Google Translate was born in 2006 as a statistical phrase-based machine translation tool and although some of the translations were laughable, there was still a use for the tool. It grew in popularity and became a mainstay Google application. In 2016, the transition to state-of-the-art neural machine translation happened and since neural networks have the ability to learn as they are used and are provided with enough data, Google Translate has developed into a powerful tool (Wu et al. 2016). The company Google has also made advances on every front in the later years and acquired many other services, like image recognition and the browser Google Chrome with Google's applications integrated making Google Translate very accessible to users. Although there is a cap of 5000 characters on text snippet translation, the cap can be raised by paying for additional services.

According to Google's blog, the transition to neural networks meant that 'It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts…' (Turovsky 2016). What is meant here is that when given an input, Google Translate browses a huge

amount of documents and resources for a sentence matching the input with an appertaining sentence in the target language.

2.3.3 DeepL

DeepL is an extension of an online dictionary called Linguee. Linguee was formed in 2009 and scraped text samples of a bilingual nature, which were then post-processed by linguists and language specialists to give additional information on the dictionary entries. Given technological advances and the rise of neural machine translation in 2016, DeepL was launched in 2017. It uses the same approach as Google Translate, uses English as a pivot language, offers text-document translation, and holds a 5000 character limit. There is also a 100.000 character limit on documents when using the free version of DeepL. A feature that separates DeepL from others is a glossary option for text translation. This is due to the foundation of Linguee's manually translated text samples, sentences, and idioms and the rest of the data available from there. This means that synonyms and similar phrasings are offered when translating live. According to their website, they outperform competitors, Google, Amazon and Microsoft, based on reviews on 119 paragraphs in different language pairs analysed by external professional translators (‘https://www.deepl.com/press.html#press_comparison_article').

Translations are offered in 24 different languages, 22 of which are European (with Portuguese offered as Brazilian as well) and two Asian, Japanese and Chinese (Simplified). There is a disclaimer though, which seems to undermine DeepL as a serious candidate for public institutions and other businesses and it reads ‘I will not use DeepL Pro for the purpose of operating critical infrastructure (as outlined in the Terms & Conditions) and acknowledge that, due to its nature, machine translation may be imprecise.’

3 Evaluation methodology

In this section, I elaborate on metrics used for the quality assessment, BLEU score, TER score and the error typology.

3.1 Evaluation metrics

3.1.1 BLEU score

To automatically evaluate on an objective basis and make machine translation evaluations more effective the Bilingual Evaluation Understudy Score was introduced in 2002 by Kishore Papineni et al. The people behind the BLEU score wanted to help ‘MT progress’ and free the ‘logjam of fruitful

research ideas' from the 'evaluation bottleneck'(Papineni et al. 2001). The idea was to perform an automatic evaluation of the quality of different machine translations cheaply and objectively. BLEU score is a metric for evaluating a machine translation by matching a candidate and a reference translation either at sentence-level, but is most suited for corpus-level evaluation. The reference translation is done by a human translator and serves as the gold standard.

The approach of the BLEU score is a combination of several computations. Fundamentally, it is a direct unigram comparison between candidates and reference sentences, thus measuring resemblance to human translations. A score is between 0 and 1, with 1 being exactly the same and 0 meaning that there is no agreement between candidate and reference sentence. As mentioned before, there are as many interpretations and translations as there are humans, so neither humans nor machines are expected to score a perfect grade. In practice, the most basic application would be to have two candidates:

1a) Candidate sentence 1: 'This initiative is has been developed to meet this challenge.'
1b) Candidate sentence 2: 'This strategy is designed to facing the dispute.'

1c) Reference sentence 1: 'This initiative has been developed to address this challenge.'
1d) Reference sentence 2: 'This initiative has been developed to meet this challenge.'

In this example, it is obvious which candidate will score the highest score for humans. Candidate sentence 1 is more intelligible and shares more words with both reference sentences. An algorithm comparing n-gram matches would also find this task easy and could effortlessly identify candidate 1 as the best translation. More n-gram matches result in a higher score. However, this is pretty naïve and not without pitfalls. The BLEU score was therefore revised and made into a modified n-gram precision score. Precision is a well-known concept in information retrieval; it is essentially a measure of how much of a selected amount of data was relevant and is accompanied by recall, which is a measure of how much relevant data was selected. To accommodate for the lack of recall in the BLEU score, a brevity penalty is introduced. The modified n-gram precision score is an implementation of a sort of inhibition of return that checks off a word in the reference sentence when it encounters a matching word in the candidate sentence. An excellent example is provided by Papineni et al. (Papineni et al. 2001) that illustrates this pitfall, namely:

‘Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.
Modified Unigram Precision = 2/7’

Had it just been the naïve unigram comparison, the score would be 7/7 because all the words in the candidate were present in Reference 1. Had the candidate sentence consisted of only two tokens ‘The the’, the score would still be disrupted. Moreover, a comparison of only unigrams would merely show the adequacy of the translation illustrated by a high score. Adequacy is a metric used for machine translations that measures a translation's understanding of the meaning. A high score on bigger n-grams, mainly 3 and 4-grams, indicates fluency, a different metric for machine translations (Snover et al. 2009). In a study on what users deemed to be worst for a translation, the researchers found disfluent texts to be the most disruptive, whereas users did not feel as strong for adequacy (Martindale and Carpuat 2018). The modified precision is thus all the checked n-gram counts for every candidate sentence in the corpus divided by the amount of candidate n-grams, which is usually no higher than 4. Machines show an exponentially lower precision when the number of n-grams is increased, which spills into how the brevity penalty ‘punishes’ shorter sentences.

As mentioned, the brevity penalty would prevent the sentence from the previous example from achieving a high score. The brevity penalty ensures that the candidate sentence corresponds to the length of the reference sentence, choice of words and word order. In cases with several reference sentences consisting of 6, 8 and 10 words respectively, a candidate sentence of eight words would correspond to the reference sentence of eight words and the brevity penalty would be 1, which is relevant for the later calculation of the BLEU score. The closest sentence would be considered the ‘best match length’. The way Papineni et at. designed this was to enforce the penalty at corpus level to avoid punishing shorter sentences and to ‘allow some freedom at sentence level’ (Papineni et al. 2001). The best-matched length sentences are summed for each candidate in the corpus, thus finding the effective length of the reference corpus. This is used for the formula:

$$r = \text{Effective corpus length}$$

$$c = \text{Candidate corpus' total length}$$

$$BP = 1 \text{ if } c > r \text{ OR } BP = e^{(1-r/c)} \text{ if } c \leq r$$

The penalty is decreasing as n-grams are increasing in correlation with the performance of the machine, namely exponentially.

It is possible to adjust the weights of n-grams to satisfy either adequacy or fluency, but for the baseline calculation of BLEU in = 4 and an equal weight between the n-grams is 0,25. Collecting all the parts of the BLEU score result in a calculation that looks like this:

$$\text{BLEU} = \text{BP} * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Here, BP is the brevity penalty, N = number of n-grams, w_n is the weight of each modified precision, (p_n), which is usually the same (0,25) for each unless different aspects of analysis are desired. This results in a measurable metric between 0 and 1, where 1 is an identical translation. (Papineni et al. 2001). Below is seen an overview of how the scaling of the BLEU score could be estimated (https://cloud.google.com/translate/automl/docs/evaluate?hl=da).

BLEU score	Quality of translation
< 0,1	Discardable translations
0,1-0,19	Almost unintelligible translations
0,2-0,29	Somewhat understandable, but with grammatical errors
0,3-0,39	Good translations
0,4-0,49	High-quality translations
0,5-0,59	Translations are of high adequacy and fluency
0,6-0,69	Equal to or better than human translations
> 0,7	Almost similar translations

Figure 1 Scale of the BLEU score

Preprocessing the data also has an effect on the outcome. Tokenisation is crucial for comparing candidate sentences and reference sentences, for instance, the removal of punctuation affects the n-grams given that commas and other punctuation marks can inflate scores.

The advantages of the BLEU score are several. Firstly, it gives a quick and measurable quantity for evaluating machine translations. Adding to this point, it can do so at a corpus-level as well with multiple references if available for better quality estimation; not only for short and simple sentences, meaning that it can be applied as a metric in other areas of natural language generation and

processing. I use the corpus-level BLEU score in my analysis. Additionally, the BLEU implementation is quite simple and straightforward and further development of it has already been completed, most notably the SacreBLEU, presented by Matt Post (Post 2018). According to the research done by Papineni et al., the BLEU score is also a useful tool, due to the high correlation it has with human judgement and a > 95 correlation coefficient of monolingual and bilingual groups is reported in their paper 'BLEU: a Method for Automatic Evaluation of Machine Translation'(Papineni et al. 2001).

3.1.2 Critiques of the Bilingual Understudy Score

The BLEU score is not without fault though and has been the subject to many raised fingers concerning several issues. As mentioned previously, human translations are considered the gold standard and there exist as many 'perfect' translations as there are translators. These may differ enough to skew the result of an evaluation and all reference translations have to meet high standards. Another issue is the lack of grammatical, syntactical or linguistic consideration. Even with a brevity penalty and modified precision, a single word makes up more than a number in the statistic. A negation, or the removal of one, has a huge influence on the meaning; a name/entity spelt wrong, like USB instead of USA, would also be considered a massive fault, but the score does not suffer adequately. This is also true when distinguishing between function words and context words. A translated text passage can usually be understood if a few function words go amiss but replace a few context words with incorrect ones and you will have a semantic disaster.

Another critique of BLEU is that there are ways to manipulate and inflate scores. The pre-processing procedure has to be the same for both candidate and reference text or the comparison is deemed invalid, as the comparison would simply not be done on the same reference. Keeping the data 'sterile' is also desirable since pre-processing has major repercussions on the scores (Post 2018). There are no standards for pre-processing in the field of machine translation, thus leaving a lot of room for divergent approaches and scoring. A trick to achieving a higher BLEU score is to get rid of obscure words if the words do not appear in a vocabulary of maybe low-resource language, thus obstructing the scores. Furthermore, the calculation of BLEU score is a conglomerate of variables that all affect the result in many ways even though there are unspoken standards, such as keeping the n-gram count to 4 and the weights evenly distributed. If you only run BLEU on unigrams, you could also change the output score. You can also negatively affect the score if you have poor grammar and were to type an error-filled sentence into a translator.

The fact that BLEU scores are calculated on sentences and strings means that it is not the overall translation quality that is analysed. The BLEU score has however been elevated by people to a guarantee of quality, despite the fact that experts call for comparisons at corpus-level instead of sentences, which would yield better results. The many reference sentences may also affect the performance of BLEU, by yielding low scores to otherwise well-translated candidates that simply do not match the references on enough criteria ('Understanding MT Quality: BLEU Scores').

BLEU does not account for anything other than how similar a translation is to one or more references and even so, synonyms are considered an error. However, there are metrics that build on BLEU that address this issue among others. METEOR (Banerjee and Lavie 2005) is an evaluation metric that stems and matches synonyms. More specifically, it does the same as the BLEU, but stems from unmatched words and tries to find a match. Furthermore, METEOR adds a penalty to the order of words. As mentioned before, there is an improved BLEU implementation called SacreBLEU that was born from a need for standardisation. Matt Post, the author of SacreBLEU, does not discard BLEU as an evaluation metric but objects to how scores are reported. The tips and tricks mentioned earlier all contribute to a murky picture when comparing BLEU scores between papers. Post (Post 2018) suggests that SacreBLEU is used instead of the regular BLEU to combat irregularities in the machine translation evaluation community and at the Workshop/Conference on Machine Translation (http://www.statmt.org/wmt21/). METEOR and SacreBLEU offer improvements to the BLEU score but are still only indicating similarity between candidate and reference sentences. They do not say anything about how much post-editing is needed, which is also a quality stamp. Translation Edit Rate (TER) is a post-editing measurement that gives you a score depending on how much effort a human translator would have to make in order to make a candidate match a reference sentence. The lower the score the better the translation (Snover et al. 2006).

3.1.3 Translation Edit Rate

Translation Edit Rate (TER) is a post-editing measurement that gives you a score depending on how much effort a human translator would have to make in order to make a candidate match a reference sentence. The lower the score the better the translation (Snover et al. 2006). A high BLEU score and a low TER score are signalling a strong machine translation. These two evaluations metrics can save a human translator time and thus create value for a company or freelance translator. TER is easy to explain as it is the quantification of editing to correct a candidate sentence. In this context,

an edit is defined as '(…) insertion, deletion, and substitution of single words as well as shifts of word sequences. A shift moves a contiguous sequence of words within the hypothesis to another location within the hypothesis.'(Snover et al. 2006). Hypothesis and candidate are synonyms in the BLEU score context. In addition, punctuation and capitalisation mistakes are treated as an edit as well. The way this enters in a TER calculation is by counting the number of edits and then dividing that by the number of word tokens in the reference sentence. Snover et al. are using more than one reference in which case you would normalise the average length of the reference sentences.

$$\text{TER} = \frac{Number\ of\ edits}{Average\ number\ of\ reference\ words}$$

However, in my case, I am only using one reference, but the calculation stays the same. In practice, this would look like this:

2a) Reference sentence: 'This Christmas, Coca Cola finally accepted blame for turning Santa red.'

2b) Candidate sentence: 'Coca Cola finally accepted blame this Christmas for turning santa red.'

Here, a few things would trigger the edit counter even though a human would be able to read and understand it easily. The shift in 'This Christmas', the deletion of the comma after the start adverbial and the missing capitalisation in 'Santa' are three edits. This gives the following formula: $\frac{3}{11}$ = 0,27=27%, which is decent. A score of 0 would be identical to a reference sentence and a 1 would require a complete rewrite of the candidate. To evaluate a full corpus, TER.corpus_score is used. It takes the total number of edits registered and divides that by the total amount of words in the corpus and then multiplies it by 100. The best TER score is 0% but is unrealistic. A good TER score is around 30% and increasing numbers means more post-editing ('https://help.inten.to/hc/en-us/articles/360020528540-MT-quality-metrics').

Like BLEU, TER correlates well with human judgement and is a good indicator of the quality of a machine translation. However, an automatic TER evaluation would need four references because 'TER score with 4 references correlates as well with a single human judgment as another human judgment does', but fewer references can work for research purposes (Snover et al. 2006). TER does not identify error types or patterns in the reference and candidate sentences and is dependent on the reference sentence being of a high standard. Mistakes in the reference sentence will inevitably be catastrophic to the whole calculation of TER. Other improvements to TER could be to adjust the weight of penalisations or add a custom vocabulary depending on your needs. For

example, insertions are to be deemed lighter than deletions. For the corpus score, you could also look to add a weighted average based on the number of tokens in a sentence. This would make a long sentence count for more than a short sentence (Snover et al. 2006).

3.1.4 Error Typology

Inspired by Larsen (2021)[6] I have created an error typology of the most common mistakes in machine translation. Potential challenges when translating to and from Danish might be compound nouns, comma rules and particle verbs (Pedersen, Rehm, and Uszkoreit 2012). The first four mistakes in bold are identified by C. Larsen and I have added some subcategories to the 'Miscellaneous' to help specify what this includes. I identified the subcategories and the need for a 'Named Entity'-category upon initial examination of the translations:

Error typology
Inconsistent terminology
Omissions/additions
Negations
Made-up words
Named entities
Miscellaneous/other than the above - Literal translation - Metaphors/Idiomatic expressions - Meaning interpretation - Lexical ambiguity - Phrasal verbs - Prepositions

Figure 2 Overview of categories and subcategories in the error typology

Inconsistent terminology: Consistent terminology is important in formal/official documents where there are a lot of technical language and terms and there can be little to no doubt about the message conveyed. If a translation tool is not able to use and adapt to domain-specific terminology, it is less applicable.

When calculating the BLEU score, synonyms are also seen as errors even though they are a big part of the general text-domain. Even though the style might be a bit looser in the general text domain, it

[6] https://sprogteknologi.dk/uploads/page_images/2022-02-23-104655.408610eTranslationpp.pdf

could still be disrupted if the jargon is not held. This could be in the finance domain, where a clean opinion means that an account gives a true and fair view and has a Danish counterpart.

Omissions/additions: This can be seen as a quality marker of how good a machine translator is at looking at context and 'understanding' meaning. A small caveat to this error type is that it can be difficult to determine, whether an omission/addition of a semantically heavy word has led to a meaning misinterpretation or the other way around. An omission would inevitably lead to a meaning misinterpretation, but it can be difficult to determine if it is the chicken or the egg. Furthermore, the addition or removal of punctuation also holds great value for a sentence. In some instances, words are simply missing in the translated version, thus distorting the meaning.

Negations: Negations hold great semantic value even though they are not deemed more important than other tokens during a BLEU evaluation. The same caveat as before is valid here since misunderstanding the sentence can lead to a negation error or that the negation error leads to misinterpretation.

Made-up words: This error type is typically seen when encountering unknown terms or new phrasings, for example during the beginning of the Covid-19 pandemic. How do the tools translate a term like Covid-19? Danish and English have different ways of spelling this.

Named entities: This is similar to the inconsistent terminology, only with names and when to translate them. There are various EU institutions and Danish ministries mentioned in the corpora and some of them have an international name, some of them do not. Furthermore, if a translation tool encounters an unknown entity, the challenge is: How to either translate it or how to find the correct version in a database.

Miscellaneous: This category seeks to hold 'the rest' and is therefore expected to be higher than the other error types. There can be sentences with correctly conjugated words and tense agreement, but with a phrasing/wording that is unintuitive in Danish or English, which can be penalized. Anglicisms or Danish phrasings also fall under this subcategory. Errors of this type can be of a semantic nature, such as meaning misinterpretation. In the general text domain, metaphors are often used to enrich text or help explain abstract concepts through non-literal methods. Metaphors and idiomatic expressions can often lead to very literal translations. These are often a challenge for a machine translator since it has most likely not seen any occurrence of such wording before and would only be able to give a good translation after encountering many occurrences of that. Another

error type that would fall under this category is the correct use of prepositions. Like negations, prepositions are important to a sentence and even though they might not have a huge impact on the BLEU score, they might hold great semantic meaning. In addition, BLEU does not distinguish between function- or context words. Lexical ambiguity describes the incorrect choice of words when a word has more than one meaning. In addition, some sentences lose their subject along the way or the word order is messed up and these are also the types of meaning interpretations that fall under this category.

I will address these errors and look at whether they are still common mistakes or have been eliminated (Larsen 2021). To get an idea of the frequency of each error and if they are still a problem, I will manually examine approximately 2% of the sentences from each corpus. Figure 4 in the appendix is and example of how I have done this.

4 Results and Analysis

In this section, I describe my data and present my results and analysis. I start by describing the corpora, how they have been processed and relevant information relating to them. Then I present the individual analysis of each tool in both English to Danish and Danish to English. I follow this with a comparison of the tools, errors and other points where they are similar and different.

4.1.1 Data: Selected corpora

To perform my analysis, several parallel corpora in different domains were needed. The European Language Resource Coordination (ELRC) consortium has a language repository for the disposal of everybody. A number of domains are available, though not all domains are supported for every language. The datasets we chose for the evaluation span several domains, all within the frame of EU interests. The parallel corpora I chose from the ELRC are from the public health, finance, cultural and general text domains (‘ELRC-SHARE’). The domains I have chosen are from a usability standpoint since these domains are deemed most useful for the end-users. In the report ‘Sprogteknologi i verdensklasse’, a big part of the participating stakeholders are involved with the public sector. I have chosen the public health and finance domains to explore the possible application of eTranslation or another tool, for companies in these sectors. In order to get a broader evaluation and to explore different challenges, I chose the culture and general text domains to counterbalance the two domains with a very strict formal writing style. If the tools are able to produce good translations for the cultural and general text domains, machine translation might be

used in new contexts, which have otherwise been restrained from using it.
The corpora are all available in an XML format. Below is an overview of the sizes of the corpora.

	Bilingual corpus from the Publications Office of the EU on the medical domain v.2	Bilingual Danish-English parallel corpus from the State Audit Office (Rigsrevisionen) website	Bilingual English-Danish parallel corpus from Aarhus 2017 – European Capital of Culture website	Bilingual English-Danish parallel corpus from Danish Working Environment Authority website
Domain	Public health	Finance	Culture	General text
No. of sentences	13242	8233	4708	1137

Table 1 Names, domain and size of selected corpora

4.1.2 Data: Description and processing

To be able to evaluate the translations given by eTranslation, Google Translate and DeepL, I will use a Bilingual Evaluation Understudy score, henceforth 'BLEU score', which will be explained in-depth in a later section. For now, it is enough to mention that it is a metric, which compares a reference sentence from the source language to a candidate sentence from the target language. This means that the corpora have to be converted from XML formatting to a .txt-file. I have a script that allows me to select and extract sentences in a language in a corpus, i.e. Danish or English, and put them in a .txt-file in a structured manner with the new line white space character '\n' added after each extracted sentence. They are ordered so each sentence has its line, which helps the sentence-matching feature of BLEU. Following this step, the corpus can be translated by a tool and used for evaluation as the gold standard. The evaluation is done by using a BLEU function from the Natural Language Toolkit: NLTK-BLEU (' https://www.nltk.org/_modules/nltk/translate/bleu_score.html'). The function itself takes a file of both the target corpora and the source corpora and matches each sentence from the reference corpus with the target corpus. However, the function does not work if the corpora do not have the same amount of sentences.

DeepL's aforementioned character cap also means that sections of the corpora have been broken into smaller files of < 5000 characters to accommodate this resulting in smaller corpora sizes. Then, DeepL translated the documents and the outputs were put into new .txt files and stitched back into one corpus to be evaluated instead of evaluating each translation snippet and averaging the scores. In contrast to DeepL, eTranslation and Google Translate do not have this cap on documents and return them in the same format they were uploaded in. On eTranslation, the translations are available for download for 24 hours and are terminated afterwards, whereas

translations from Google Translate are downloaded immediately. When performing translation tasks with the three tools, the structure of the files, e.g. line breaks, sections and so on, are maintained in the output, which eases post-editing.

The three translation tools can be assessed and compared on different parameters and across different domains. Firstly, I will look at the individual performance of eTranslation in both English to Danish and Danish to English and then do the same for Google Translate and lastly for DeepL. Secondly, I will highlight different areas that distinguish the three tools from each other and compare their performances. I look at the BLEU scores, TER, and analyse manually registered error types in extracts of 2% of the sentences from the corpora. Even though eTranslation offers domain-specific translations, all corpora were translated using the general domain to ensure the fairest foundation for comparison between the tools.

4.2 Individual assessment

Overall, the performances of eTranslation, Google Translate and DeepL are very good. All the translations are readable and fluent. It is therefore easier to highlight the mistakes to illustrate the shortcomings and challenges there are, instead of finding examples of successful translation, which are plenty. It is worth noting that a successful translation can be a candidate sentence that is identical to a reference sentence when measuring BLEU and TER. I have colour coded the results of the BLEU and TER scores to help readability. I change colour every 0,1 decimal and 10% point.

The error typology tries to find readable and well-translated sentences, not necessarily the most identical like BLEU. This means that a successfully translated sentence in this instance is not necessarily a complete copy of the reference sentence, but is a sentence that is readable and captures the correct semantic meaning.

4.2.1 Results for eTranslation:

Unsurprisingly, the BLEU score suggests that eTranslation has done well on the datasets that are available to it through ELRC. The general impression when examining the translations is that they are fluent and readable, which is reflected in the > 0.65 scores for both English to Danish translations and vice versa (https://cloud.google.com/translate/automl/docs/evaluate?hl=da). This high standard is without any smoothing applied and the lowest scoring translation is 0.679 seen in the Danish to English unsmoothed finance domain translation in table 1. Although the general text domain scores significantly lower, it will also be treated as a chapter for itself due to its difference from the other domains. The common interpretation of the score range is that everything above 0.60

could be considered competitive with human translations. Even human translators do not score a 1 on other human translations since they rarely produce identical translations. A score between 0.30 and 0.40 would be considered a comprehensible translation without grave grammatical mistakes (https://cloud.google.com/translate/automl/docs/evaluate?hl=da).

English to Danish	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,818	0,728	0,685	0,706
Smoothed	0,909	0,758	0,775	0,737
TER	35,33%	44,99%	60,95%	53,83%

Table 2 eTranslation's BLEU and TER-scores for English to Danish translation

The English to Danish scores are very good for both evaluation metrics. The high BLEU scores mean the translations are very close to the reference corpus. Furthermore, the low TER scores suggest that some-to-little post-editing would be required to produce a similar sentence to the reference.

One of the differences between the domains is the way semicolons are handled by eTranslation. Semicolons are rarely used in Danish in comparison to English due to the strong Danish comma. It would be reasonable to think that a comma would replace a semicolon in these translations from English to Danish or a full-stop leading into a new sentence. However, in the finance corpus, all 51 semicolons in the finance corpus are replaced by full stops, the following word is not capitalised, which leads to errors since the semicolons occur mid-sentence as seen in the example below, highlighted in bold.

3a) Source sentence: 'The objective of the Danish national parks is not only to strengthen and develop nature**;** national parks must also meet other objectives, eg, to promote an understanding of nature, tourism and business development.'

3b) Candidate sentence: 'Formålet med de danske nationalparker er ikke kun at styrke og udvikle natur. de nationale parker skal også opfylde andre mål, f.eks. at fremme forståelsen af naturen, turismen og erhvervsudviklingen.'

In the public health and general text corpora, semicolons are simply omitted in the translation without any replacement punctuation and finally, in the culture corpus, semicolons are kept in the Danish translation. The public health BLEU score is very high and is almost a 1:1 copy of the reference. This is also reflected in the TER score, which is low without being as impressive

as the BLEU score for the public health translation. There is little post-editing to be done on this corpus, but the same can not be said about the finance domain translation with a score of 60,95% meaning that the amount of edits is over sixty percent of the number of words.

Below is a table showing the amount of error type occurrences registered in extracts of the corpora. The error types are registered across the whole extract and a sentence can contain more than one occurrence of the same error type. The number in parenthesis is the number of sentences an extract of text consists of.

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	4	2	0	0	37	0
Culture (82)	3	0	0	0	36	6
Finance (164)	11	0	0	0	36	0
General (23)	10	0	0	0	4	2

Table 3 Error types for the English to Danish translations from eTranslation

Given that the content of the four corpora is EU-related, you would not expect EU-related terminology and named entities to be an issue. This is also the case as table 2 shows as there are low amounts of errors and a majority of the translated sentences convey the same meaning as the reference. The most noticeable thing in this table is the few-to-non-existing errors other than inconsistent terminology and miscellaneous. The inconsistent terminology is mainly a reoccurring single term, like ‘near miss’ in the finance corpus, which in this context is a term related to the Danish Working Authority. In the Danish reference corpus, the correct term is ‘nærved hændelse’, but it has been translated in a plethora of different ways like ‘næsten fejl’,’nær ved fejl’, even ‘nær-miss’ or ‘nær miss’. Besides the incorrect translation, the incoherent translation is interesting to note as well, since it is the same wording to be translated throughout the corpus.

An error in the miscellaneous category in the culture corpus highlights the aforementioned difficulties with idioms and underlying meaning. This is an example of eTranslation’s attempt at handling an idiomatic saying:

4a) Source sentence: ‘Thus, we also highly value the good advice and recommendations we've received 'along the way',” says CEO of Aarhus 2017, Rebecca Matthews.’

4b) Candidate sentence: ‘Derfor sætter vi også stor pris på de gode råd og anbefalinger, vi har modtaget "på lang vej'," siger CEO for Aarhus 2017, Rebecca Matthews.’

The fact that ‘along the way’ is in quotation marks indicates some kind of deeper understanding is needed. The translation is a blatant mistake that underlines the difficulties eTranslation still faces. Another thing I note is the translation of ‘CEO’ or the lack thereof. CEO is used in Danish, but has an international/business profile to it, whereas the Danish word ‘Direktør’ is readily available. It is difficult to know when and what to translate and what to leave. One of the few named entity errors is an example of exactly that. In the culture corpus, the report mentioned in this sentence ‘This is all to be found in the report ”Aarhus European Capital of Culture 2017 - Second Monitoring Meeting” that the EU has only just published on their homepage.’ has been translated into ‘Aarhus Europæisk Kulturhovedstad 2017 — Andet overvågningsmøde’, which is not an incorrect translation. However, when looking at the Danish reference corpus, the name of the report is not translated and the correct thing would be to leave the title as it is.

An example of how eTranslation does everything right, but still falls short is the translation of this sentence:

5a) Source sentence: ‘In cases where notification must be made immediately to the South Jutland Police or the Danish WEA pursuant to section 8 in the Executive Order on Notification, the operator and the owner, respectively, have the duty of notification.’

5b) Candidate sentence: ‘I tilfælde, hvor anmeldelse skal ske straks til Sønderjyllands Politi eller Arbejdstilsynet i henhold til § 8 i bekendtgørelse om anmeldelse, har henholdsvis operatøren og ejeren underretningspligt.’

I want to highlight that South Jutland Police is the correct English name for ‘Syd- og Sønderjyllands politi’, but as is evident in the Danish name the police force covers the Danish region ‘Sønderjylland’ and the southern part of Jutland, hence the name needs to have 2 mentioning of south. eTranslation has translated this into ‘Sønderjyllands Politi’, which the correct literal translation of the input given. However, it is actually not the correct translation of the entity, but how should it know?

In general, eTranslation does well on English to Danish translations and even better on Danish to English translations, though only by a few decimal points. Looking at table 3, it is evident that they are performing at a similar level.

Danish to English	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,821	0,743	0,679	0,707
Smoothed	0,925	0,772	0,828	0,741
TER	33,48%	44,94%	58,18%	49,58%

Table 4 eTranslation's BLEU scores for Danish to English translation

The BLEU scores follow the same trend as before with public health at the top and finance at the bottom, though still a respectable score. The nature of the corrections in Danish to English translations is much the same as the English to Danish translations. The high BLEU scores indicate very well translated corpora where the mistakes are mostly lexical choices, though the translated sentences still convey the correct meaning.

Like the BLEU score, the TER scores are a mirror image of the English to Danish analysis. It is worth noting that the translation of the finance corpus is improved and now below sixty percent. The public health TER score is low again and the two other fall somewhere in between the two extremes and corresponds well to their BLEU scores.

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	3	0	0	0	24	5
Culture (82)	0	0	0	0	28	0
Finance (164)	16	0	0	0	43	0
General (23)	9	0	1	0	7	5

Table 5 Error types for the Danish to English translations from eTranslation

The most striking thing when looking at table 4 is the columns of zeroes. There are no examples of omission/addition errors or negation errors whatsoever. Even the made-up word is an entity, meaning it could belong in the named entity category. It is definitely a positive that these sorts of errors are non-existing and proves that eTranslation can be used to produce high-quality translations although with some reservations.

In the finance corpus, there is an example of the difficulties when looking at domain-specific terms. The Danish sentence to be translated looks like this: '(…)skal afrapporteres i form af en revisionspåtegning og revisionsberetning senest 15. maj i året efter regnskabsåret,(…)'. The

correct English translation for the two financial terms ’revisionspåtegning’ and ’revisionsberetning’ are ‘auditor's opinion’ and ‘auditor's report’, but have been translated into: ‘(…)must be reported in the form of an audit report and audit report no later than 15 May of the year following the financial year(…)’. There is no distinction between the two terms and the sentence loses some of its meaning.

The only instance of an error with made-up words within this language pair using eTranslation is in the general text-domain. However, this could also be a named entity error. The name to be translated is ‘Arbejdstilsynet’, which has been translated correctly in the English candidate to ‘the Working Environment Authority’. In the same candidate translation, another occurrence of ‘Arbejdstilsynet’ has been translated into ‘the Labour Inspectorate’, which is not incorrect, as other countries' counterpart is called Labour Inspectorate. eTranslation’s Danish to English translation also struggled with the police force of Southern Jutland. The resulting error is the same literal translation of ‘Syd- og Sønderjyllands politi’ into ‘South and South Jutland Police’, but as mentioned before, the English name is South Jutland Police.

Table 5 sums up and compares the quality between the two with the highest score within each domain marked by bold. As is evident, Danish to English translations are superior in all but one instance.

EN-DA / **DA-EN**	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,818 / **0,821**	0,728 / **0,743**	**0,685** / 0,679	0,706 / **0,707**
Smoothed	0,909 / **0,925**	0,758 / **0,772**	0,775 / **0,828**	0,737 / **0,741**
TER (in %)	35,33/**33,48**	44,99/**44,94**	60,95/**58,18**	53,83/**49,58**

Table 6 Comparison of eTranslation’s BLEU scores

There is not much between the translations from the two language pairs, where the highest-scoring pair is Danish to English. Both in terms of BLEU score quality and the nature of errors. From the scores reported, eTranslation can definitely be used as a tool for translation tasks for the stakeholders with access to it. Even the lowest scoring translations are still of adequate quality, but should not be used without post-editing as illustrated by the error typology. eTranslation also offers domain-specific translations, but actually scores lower than when using the general text function. These scores can be found in the appendix in tables 18 and 19. Examining the domain-specific

translations indicated a reduction of a general vocabulary, but a better domain-specific vocabulary for the domain in question.

The low editing numbers are further backing the usefulness of neural networks. As Larsen (2019) concludes in his rapport 'Neural Machine Translation in DA Brief Assessment Report' on why a transition to neural machine translation is beneficial that '*the DA LD is overwhelmingly in favour of NMT (94 % of respondents prefer NMT against only 6% SMT). Although the quality is far from what is found in human translation, respondents appreciate its improvement over SMT because less editing is needed.*' The issue when this report was conducted was the otherwise low quality that the language pair had. The generally low quality of machine translations is also pointed out, but as mentioned previously the quality has surged upwards in recent years and you can see the benefit from it in the scores reported.
The errors I found are not completely ruining the translated sentences and it is a mostly stylistic and deeper semantic type of error.

A tool that does not have any domain-specific setting is Google Translate.

4.2.2 Results for Google Translate

Google Translate's BLEU score results are of a similar quality to what eTranslation performed. I will be using the same error type backdrop as on eTranslation to help identify what sort of errors are present in the translations from Google.

English to Danish	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,780	0,757	0,709	0,722
Smoothed	0,876	0,784	0,798	0,752
TER	40,97%	43,79%	58,13%	51,98%

Table 7 BLEU score for Google Translate on English to Danish translations

The absolute highest score is within the public health domain with a score of 0.876 when smoothed. The rest of the scores are similar and follow the same trend as eTranslation, with public health translations at the top and finance at the bottom. Given that the datasets are available online to stakeholders, such as Google, it is not surprising that Google Translate scores are very high.

The characteristics of Google's translations from English to Danish are Anglicisms like in the following example from the finance corpus.

6a) Source sentence: ‘Unless Denmark withdraws from the programme, a Joint Strike Fighter acquisition will be exempt from the ordinary rules(…)’

6b) Reference sentence: ‘En anskaffelse af Joint Strike Fighter vil, medmindre Danmark udtræder af programmet, være undtaget fra de almindelige regler(…)’

6c) Candidate sentence: ‘Medmindre Danmark udtræder af programmet, vil et Joint Strike Fighter-anskaffelse være undtaget fra de almindelige regler(…)’.

This way of constructing a sentence is not wrong, but depending on what information you want to convey, you would write it differently in the two languages. English sentences are generally constructed with end-weight meaning that the most important information is at the end of a sentence. However, this is not the case for Danish, as the end of the sentence is less important than in English. The Danish reference for the example sentence above is constructed in a way that sounds way more Danish.

The TER score reveals difficulties with the finance corpus especially, but also moderately high TER scores for the rest of the domains. Google Translate does relatively well on the English to Danish translation except for the finance.

The types of errors are more interesting as evident in the table below.

Error types	Inconsistent terminology	Omissions / additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	8	3	0	0	46	5
Culture (82)	0	0	0	0	21	13
Finance (164)	23	10	0	0	52	4
General (23)	8	2	0	0	4	4

Table 8 Error types for the English to Danish translations from Google Translate

The first thing to notice is the low amount of errors in general. The culture domain has 4 columns of zero. This indicates a vast majority of well-translated sentences. The two ‘empty’ columns of made-up words and negation errors show that Google Translate has come a long way. Furthermore, the fact that there are no errors regarding made-up words shows that Google Translate is either very resourceful when it comes to locating and combining the right words or simply trained way better

and is learning the correct lexical pair very well. In addition, the low number of incorrectly named entities shows that Google Translate is capable of using its search engine or translation memory to locate the correct name.

When taking a closer look at inconsistent terminology, Google Translate mostly struggles in the finance domain and it is the same issue as seen in eTranslation's Danish to English. The two audit statements in the source sentence below have been translated into the same word and look like this:

7a) Source sentence: 'A completed annual audit of an institution, (…) must be reported on by means of an auditor's opinion and an auditor's report(…)'

7b) Candidate sentence: 'En gennemført årlig revision af en institution,(…) skal aflægges rapport om ved hjælp af en revisionspåtegning og en revisionspåtegning(…)'.

Another mistake, which is on the border of two categories, omission and miscellaneous respectively, is a missing full stop after the abbreviation 'mv' in this sentence from the finance domain '(…)mellem lande, sektorer, bistandsinstrumenter mv(…)'. However, there is also an occurrence of a successful translation, where the abbreviation is followed by a full stop, as it should be. This does confirm that it is an omission error and not a lack of knowledge of Danish abbreviation rules. Google Translate also shows signs of Anglicism in the finance corpus. The following title of the national park is correctly translated, however, the English possessive apostrophe and s is kept: 'Thy National Park"s' into 'Nationalpark Thy"s'. I have also noted a weird phrasing in the translation of the public health corpus. The English sentence is: '(a) the contamination lasts for more than four consecutive months; or' and the translation is this: '(a) forureningen varer i mere end fire på hinanden følgende måneder eller'. The phrasing 'på hinanden følgende måneder' can be literally translated to 'on each other following months', which is an unintuitive way of phrasing 'consecutive' for both languages.

Google Translate does well on named entities and the errors in this category do not suggest a lack of knowledge, but execution. I found that Google Translate suffers the same problem as eTranslation. In the culture corpus, the report 'Aarhus European Capital of Culture 2017 – Second Monitoring Meeting' has not only been translated, as it should not but also translated to different versions like 'Aarhus Europæisk Kulturhovedstad 2017 - Andet Overvågningsmøde' and 'Aarhus Europæisk Kulturhovedstad 2017 - Second Monitoring Meeting'.

Google Translate's English to Danish translations are very good across all the domains and would be suitable for many purposes. It is the same story with the Danish to English translations.

Danish to English	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,788	0,759	0,684	0,718
Smoothed	0,909	0,786	0,833	0,752
TER	41,21%	43,60%	58,28%	48,46%

Table 9 Google Translate's BLEU score for Danish to English translations

The Danish to English results are showing the same as the others, where the public health corpus translation scores the highest score and the other domains are in the same places when unsmoothed. Here the smoothed score of 0.909 is one of the absolute top scores, but similarly, the other scores make for great translations as well. The errors in the Danish to English translations are also mostly identical to the English to Danish.

One observation to be noted is the loss of the style format and Danish writing style. I would place the following example under inconsistent terminology. This is from the finance corpus:

8a) Source sentence: 'According to section 1 of the act its objective is to contribute to sustainable development of the rural areas(…)'

8b) Candidate sentence: 'Formålet med loven er ifølge § 1 at bidrage til en bæredygtig udvikling af landdistrikterne(…)'

8c) Source sentence: 'According to section 1 of the act its objective is to contribute to sustainable development of the rural areas(…)',

8d) Candidate sentence: 'Ifølge lovens § 1 er formålet at bidrage til en bæredygtig udvikling af landdistrikterne(…)'

Due to the alphabetically ordered corpus, the sentence is accompanied by other sentences with the same initial wording of 'Formålet med…. which shows that somewhere either at Rigsrevisionen or the creation of the corpus the Danish wording was lost. This results in a Danish candidate sentence that has broken the stylesheet and looks like sentence 8d. The translation itself is correct but the sentence does not fit in and is very easily identified in the candidate corpus, as the translations are returned in the same order as they are iterated. It is nestled between other sentences with the exact same start to the sentence, namely 'Formålet med…'.

Furthermore, Google Translate struggles to choose one English formulation for the phrase 'Formålet med', as it alternates between 'the objective of the(se)' and 'the purpose of the(se)' where the only identifiable determiner is what noun immediately follows. Of 38 sentences with this wording or the similar 'Formålet er…', 18 began with 'The purpose', 13 with 'The objective' and a single sentence began with 'The aim is to'. The reason I include this latter example is that sentences beginning with 'Formålet er…' also show this differing pattern and are in the same error group and the verb would therefore not be the reason. You would expect that such phrasing would be consistent within a domain with a formal tone and glossary.

The TER scores tell the same story as before with a finance corpus that causes issues. Interestingly, the general text domain has a score below fifty percent and the culture and public health domains are almost similar.

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	18	22	0	1	31	8
Culture (82)	4	0	0	0	15	0
Finance (164)	27	16	0	0	71	16
General (23)	5	0	0	0	6	2

Table 10 Error types for the Danish to English translations from Google Translate

Once again, the finance corpus has the most errors and the culture domain has very few errors and four columns of zeros. There are recurring errors, like the difficulties with distinguishing technical terms in the finance corpus, where 'revisionspåtegning' and 'reviosionsberetning' were translated into 'auditor's report'. There are other difficulties with the finance vocabulary, as the term 'Modkøb' has been translated into the literal 'counter-purchase'. This is not the correct translation, however, it is not an error that disturbs the meaning of the sentence. It is positive that it is these sorts of errors I find and not made-up nonsensical attempts at translating a technical term. Although, this translation in particular is a bit off from the correct translation, which is 'offset obligation'.

A trait of the public health corpus translation is the insertion of full stops mid-sentence and not having the following word capitalised. It is evident in the following sentence, which is a candidate sentence from the public health domain: '(…)as a result of the covid-19 outbreak and is

directly linked to its creation. or the extension of reduced working time schemes(…)'. This issue and seemingly random insertions of semicolons at the end of clauses or sentences in the corpus that otherwise would end without any punctuation are what make up the biggest part of the omission/addition errors of the public health domain.

The only case I have found of a possible negation error is in the public health corpus translation, but given that this type of error has not been present in ANY of the other translations, it is worth taking a closer look. The translation looks like this:

9a) Source sentence: 'Der kan højst ydes godtgørelse i henhold til stk. 1, litra a), i højst 12 måneder(…)'

9b) Candidate sentence: 'Reimbursement may not be granted in accordance with subsection (1). 1 (a) for a maximum of 12 months(…)'

The error is obvious and the insertion of a negation completely changes the meaning of the sentence. This is a case of the chicken and the egg, since the insertion leads to a meaning misinterpretation, but can also be the other way around, where a meaning misinterpretation led to the insertion. I have classified it as a negation error, but it is the sole example of such a mistake.

Google Translate also show some struggles when dealing with lesser-known words. In the culture domain, a citation from the CEO illustrates just that, as the statement in Danish goes: 'Det er meget vigtigt for os, at vi får sådan et flot skudsmål.' and the incorrect translation produced by Google Translate goes: 'It is very important for us that we get such a great shot.' The Danish word 'skudsmål' contains 'skud' meaning shot and 'mål' meaning goal, but the definition of 'skudsmål' is an assessment or rating, which is certainly not reflected in the translation 'a great shot.'

In the named entity category, Google Translate does not do well with the 'Syd- og Sønderjyllands politi' as it translates it into **'**South and South Jutland Police', which has already been mentioned. A more straightforward error is how 'Miljøministeriet' has been translated into 'the Ministry of the Environment' in the finance corpus translation. When conferring with the English version of the Ministry of Environment of Denmark[7] a clear answer is given to what a correct translation would be. You could excuse this error, as it is actually wrongfully stated in the English reference corpus as

[7] https://en.mim.dk/

'the Ministry of the Environment'. This shows that a translation service is reliant/vulnerable to dependable data and that Google Translate was not able to retrieve that information.

EN-DA/**EN-DA**	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,780/**0,788**	0,757/**0,759**	**0,709**/0,684	**0,722**/0,718
Smoothed	0,876/**0,909**	0,784/**0,786**	0,798/**0,833**	0,752/0,752
TER (in %)	**40,97**/41,21	43,79/**43,60**	**58,13**/58,28	51,98/**48,46**

Table 11 Comparison of Google Translate's BLEU and TER-scores

Overall, Google Translate is a very good translation tool and would all be suitable for use in many cases. It comes as no surprise since Google Translate could possibly have trained their algorithm on these datasets or at least retrieved them. The error types are very similar in the two language pairs as can be seen in the technical term issues and the repetitions of the same named entity problems.

The TER scores once again reveal that the finance corpus is more tricky than the other domains despite being the BLEU scores being close to general text. All the TER scores suggest that there is a moderate amount of post-editing to do if you use Google Translate.

A tool claiming to be four times as good as Google Translate is DeepL.

4.2.3 Results for DeepL

As mentioned, I have used the free version of DeepL, which has a limit of 5000 characters pr. translation. I have therefore translated snippets of 5000 characters and patched them together into corpora of around 50.000 characters, which serves as the basis for quality estimation.

DeepL performs well when translating the text snippets, as is visible in the tables in this section. The lowest scoring domain for DeepL is the public health domain with a very respectable 0,621 and the culture domain scoring the highest BLEU score with 0,770. Interestingly, the order is different from the other tools.

English to Danish	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,621	0,770	0,708	0,719
Smoothed	0,759	0,906	0,823	0,834
TER	68,55%	42,40%	57,33%	53,14%

Table 12 DeepL's BLEU scores for English to Danish translations

DeepL produces very convincing translations, but even with the high BLEU scores, there are still some issues with the English to Danish translations. In the culture domain, DeepL demonstrates little knowledge of Danish writing style and format. The following examples do not result in unreadable translations but illustrate different areas of improvement for the English to Danish translation in general. The first is about possession. As seen previously, the English possessive construction of apostrophe s has snuck its way into this candidate sentence: '(…)den stærke regionale forankring og Aarhus 2017's strategiske forretningsplan(…)'. The correct Danish way to express genitive is to put an –s on the end of a word and omit the apostrophe unless the word ends in –s, -x or –z. This is also true for abbreviations and numbers and symbols. The second issue is compound nouns and when to use them. The sentence: '(…)the panel also emphasizes a clear organizational structure and project management.' has been translated into '(…)fremhæver panelet også en klar organisatorisk struktur og projektledelse.' Here, only one of the potential two nouns has been brought together into one word, however, I have found no clear strategy as to when a compound noun is created and when DeepL leaves it as two words. In the Danish reference corpus, the sentence looks like this: '(…)fremhæver panelet også en klar organisationsstruktur og projektledelse.'

The 'low' BLEU score for the public health domain is primarily a result of small deviations in the choice of words and phrasing like in this example:

10a) Reference sentence: 'De specifikke EHFF-foranstaltninger suppleres med en ændring af forordningen om den fælles markedsordning for at:'
10b) Candidate sentence: 'De specifikke EHFF-foranstaltninger suppleres af en ændring af forordningen om den fælles markedsordning med henblik på:'

The TER scores are very high for the public health domain translation and are an indication of a below-average translation. This alongside the two domains scoring above fifty tells a story that the translations might not be as good as the BLEU score suggests.

This is also reflected in the error typology for the English to Danish DeepL translation below.

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	0	4	0	0	32	10
Culture (82)	4	3	0	0	23	5
Finance (164)	36	8	0	0	63	8
General (45)	1	0	0	0	8	5

Table 13 Error types for the English to Danish translations from DeepL

DeepL does not have issues with negations and can also handle new/unknown words well enough to avoid making up words. In addition, inconsistent terminology only seems to be a problem in the finance domain, as the public health translation has zero occurrences of inconsistent terminology. The biggest issue for DeepL is entities as displayed in Table 13.

One of the many examples of inconsistent terminology in the finance domain translation is the translation of 'GNI-based resources' into 'BNI-baserede indtægter.' Looking at the Danish reference, the term in question is 'BNI-bidrag', which translates to 'GNI-contributions'. This is also an error in the omission category, as some vital information has been omitted in the Danish translation since 'based' has been left out. Interestingly, DeepL has successfully translated the 'auditor's opinion and auditor's report' into two distinct terms, 'revisorerklæring og en revisionsberetning'.

As mentioned before, DeepL does struggle with the genitive, which has resulted in insertions of apostrophes in the Danish candidates. These sort of errors are what makes up most of the additions in the finance and culture domains. However, in the public health domain, the nature of insertions is random numbers as seen in this translation. The sentence 'Contributions referred to in paragraph 1(…)' is translated into '2. De bidrag, der er omhandlet i stk. 1(…)', which can not be explained by any meaning misinterpretation, but simply as an error in the output.

Although DeepL produces good translations, there are signs that there is no real understanding of what is and should be translated. In the culture domain, the sentence 'Now everyone wants to "gentænke"' 'gentænke' is established as a concept. In the following sentences, 'gentænke' is mentioned several times, but in different ways:

11a) Source sentence: '(…)applying the concept of rethinking.'
11b) Candidate sentence: '(…) de anvendte begrebet gentænkning.'

12a) Source sentence: '(…) the word 'rethink' has now fought its way to (…)'
12b) Candidate sentence: '(…) har ordet "rethink" nu kæmpet sig ind (…)'

13a) Source sentence: '"Gentænk" has become a frequently used word(…)'
13b) Candidate sentence: '"Gentænk" er blevet et hyppigt brugt ord(…)'

DeepL falls victim to inconsistency in the source corpus, since DeepL can only translate the input DeepL gets, thus resulting in the varying translation above. This might sound obvious, but the point is that DeepL does not have the ability to 'remember' and make the anaphoric reference to the earlier established concept' gentænke'. It seems that translation happens depending on the quotation marks since the word in single quotation marks has been translated. Whether a concept like 'gentænke' should be translated or not is difficult to judge since it is not a culturally laden word nor is it inherently Danish, as a concept like 'Hygge', but in this example, it should be consistently translated to either of the languages.

The named entity errors in the general text domain are simply untranslated occurrences of the abbreviated form of the Danish Working Environment Authority, WEA as seen in this example:

14a) Source sentence: 'In this case, the Danish WEA will contact the operator and the owner, respectively(…)'
14b) Candidate sentence: 'I dette tilfælde vil den danske WEA kontakte henholdsvis(…)'

Overall, the English to Danish translations are very readable with the majority of the errors being of semantic nature. It is much the same story for DeepL's Danish to English translations.

Danish to English	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	0,615	0,774	0,692	0,690
Smoothed	0,751	0,900	0,795	0,806
TER	68,82%	41,03%	54,56%	50,43%

Table 14 DeepL's BLEU scores for Danish to English translations

The standard of Danish to English translations from DeepL are very high. There are no issues with style or format similar to the English to Danish translations, which points towards mismatch in words and phrasing as the main reason behind the 'low' BLEU scores. Once again, the public health domain score is the lowest with 'only' 0,615 and the culture domain is convincingly at the top with 0,774. Opposite to the English to Danish, the finance domain score is higher than the general text domain, but only by a very small margin. This is not reflected in the TER scores as the general text domain scores lower than the finance translation and you could thus argue that the general text translation is better than the finance. They are both subject to a good part of post-editing as they both score above fifty percent. The public health domain translation is not useless, but would require a lot of editing as almost seventy percent of the words would be subject to an edit of the kind mentioned in the TER section.

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	0	22	0	0	68	0
Culture (82)	6	5	0	0	24	8
Finance (164)	25	13	0	0	65	12
General (45)	10	3	0	0	11	0

Table 15 Error types for the Danish to English translations from DeepL

In the Danish to English translation, there are fewer named entity errors than in the English to Danish translation. Furthermore, the four columns of zeros in the public health domain support the claim that the translations actually are very good, even though the BLEU scores are lower than 0,7. One of the many terminology errors in the finance domain is the inability to correctly translate a certain type of VAT fraud, namely VAT carousel fraud.

15a) Source sentence: 'Forholdene kan være indikationer på momskarruselsvindel.'
15b) Candidate sentence: The circumstances may be indicative of VAT fraud.'

The correct translation of this type of fraud is the very literal 'VAT carousel fraud', which DeepL fails to get in the Danish to English translation, but has done correctly in the English to Danish translation. Another mistake that seems easy to avoid is an omission in the finance domain. The end of the sentence is simply left out.

16a) Source sentence: ‘En bedre koordinering vil således kunne imødegå unødigt dobbeltarbejde.’
16b) Candidate sentence: ‘Better coordination could thus avoid unnecessary duplication.’

DeepL has somewhat successfully dealt with the concept of rethinking in the culture corpus. In the Danish to English translation, the concept is consistently translated to the same word, except for one instance, where it is simply the wrong choice of words and ‘gentleness’ is the word in question instead of ‘rethink’.

17a) Source sentence: ‘I mediestrømmen er gentænk blevet et hyppigt anvendt ord (…)’
17b) Candidate sentence: ’In the media stream, gentleness has become a frequently used word (…)’

An error that is more difficult to correct is the choice of the preposition to accompany ‘important’. Depending on whether you want to express that the important thing is personal to the subjective (important to) or that the important thing is used to accomplish something (important for), like a goal in football is important for winning the match. It can be difficult to select which is the correct option for this example. In the reference sentence, it is deemed that the assessment is important to something implied in the context, whereas the candidate suggests that subject currently has achieved something important.

18a) Reference sentence: ’It is very important for us to receive such an impressive assessment.’
18b) Source sentence: ‘Det er meget vigtigt for os, at vi får sådan et godt skudsmål.’
18c) Candidate sentence: ’It is very important to us that we get such a nice shot.’

Furthermore, this example also exposes issues with idiomatic language, as seen in the translation of ‘skudsmål’.

EN-DA/**DA-EN**	**Public Health**	**Culture**	**Finance**	**General Text**
w/o smoothing	**0,621**/0,615	0,770/**0,774**	**0,708**/0,692	**0,719**/0,690
Smoothed	**0,759**/0,751	**0,906**/0,900	**0,823**/0,795	**0,834**/0,806
TER (in %)	**68,55**/68,82	42,40/**41,03**	57,33/**54,56**	53,14/**50,43**

Table 16 Comparison of DeepL's BLEU scores

Table 16 illustrates that DeepL performs best on translation from English to Danish, though only by a few points. Although there is a higher frequency of errors in the English to Danish translations, the BLEU scores suggest the production of a more similar translation to the reference corpus. This

is because of the inadequate knowledge of the Danish language. DeepL falls short when the issues are decisions on style and format and not just straightforward translations. In general, DeepL is a good tool for texts in these domains.

The TER scores for DeepL are not encouraging and a lot of post-editing has to be done especially for the public health domain in both languages. It is interesting that the Danish to English general text and finance translations score lower than the English to Danish translations, which have scored a higher BLEU score.

4.3 Comparative assessment

In this comparison, I will compare the evaluation scores for the tools. In addition, I will look at general points of improvement for the tools, highlight areas, and error types that separate them from each other.

It is evident that there is very little between the tools regarding the quality of translations they are able to produce. eTranslation was expected to do well due to the nature of the corpora, Google Translate has been the most used tool and therefore the expectations are high and finally DeepL claims to be 4 times better than Google and other market competitors. It is interesting to highlight areas of improvement as the tools do vary in error types and now look into new challenges. It is also evident that machine translation has moved past the point where negations and made-up words are serious issues. Furthermore, there were very few to no syntactical errors, subject-verb agreement mistakes, singular and plural nouns were in agreement and pronouns referred back to the correct subject.

4.3.1 The metrics scores

The three translation tools all score very high BLEU scores. Furthermore, the scores are all very similar across the domains except for DeepL's public health scores. Still, the lowest score of 0,615 from DeepL's Danish to English public health translation indicates a strong translation. A possible explanation as to why the scores are that high is that the domains all have a closed vocabulary. There is nothing besides DeepL's public health BLEU score that can significantly separate them, which is around 0,2 lower than eTranslation.
Remarkable that DeepL's scores are low for the public health translations in both languages relative to the two other machines. This point will be elaborated on in section 5. Certainly, DeepL is not 4 times better as proclaimed on their website, but it definitely produces translations on a similar level to the two other tools. The BLEU scores suggest that translations from English to Danish and from

Danish to English have progressed to a point, where they can be qualified as above-average/high standard using these three tools.

The TER scores reveal that DeepL have some flaws and using this tool will result in more post-editing compared to the other tools. eTranslation's TER scores are the lowest among the tools and Google Translate slots into the middle. It was remarkable that some translations broke the inverse proportionality between the BLEU and TER score, which dictates that when BLEU is high TER must be low. This happens in DeepL's general text translations, where the translation with the lowest BLEU score also scores the lowest TER score and vice versa for the highest scoring. The TER score contributes to a better quality estimation of the three tools.

The tools:

An overall decision when translating any kind of language into English from DeepL is what kind of English to translate into, British English or American English. eTranslation and Google Translate do not have this option yet or have simply chosen not to, however, the tools perfectly well understand American English, but do not have an option to output American English. Interestingly, the output from Google Translate is American English, as words like organisation, analyse and characterise are spelt with a z instead of an s. The standard output from eTranslation is British English and I chose British English for the DeepL translations, which DeepL have been well-executed.

4.3.2 Error types

Inconsistent terminology:

A recurring problem was the terminology of the finance corpus, where the most problematic part was the auditor's report/revisionspåtegning terms. The tricky part for the tools is to distinguish between two closely related technical terms. Only DeepL managed to produce correct translations for both languages, whereas eTranslation did not translate correctly into English from Danish and Google Translate failed both ways. However, DeepL produced the only failed translation of VAT carousel fraud. Both Google Translate and eTranslation correctly translated the term from English to Danish and vice versa, but DeepL only did so for English to Danish. The Danish to English translation does not specify what kind of VAT fraud is mentioned, which results in a big loss of information.

In the error typology analysis of the public health domain, I found that the term 'Covid-19' is treated differently across the tools, but also across languages. The Danish dictionary suggests two

correct spellings, covid-19 and COVID-19, whereas Cambridge Dictionary suggests Covid-19 and COVID-19 as valid spellings. eTranslation prevailed when correctly translating the terms into the respective languages, whereas DeepL only succeeded with the English to Danish where it is spelt 'COVID-19'. However, DeepL's Danish to English translation failed due to occurrences of missed hyphens. Google Translate also only produced one good version, which is the English to Danish where 'COVID-19' is consistently used. Google Translate's Danish to English translation resulted in the term 'covid-19', which is not the correct translation according to the Cambridge Dictionary.

Insertions:

The majority of insertions and deletions were in the context of EU legislations, regulations, and articles, which there are a high frequency of in these particular corpora. The characteristic of an article mentioned in a reference corpus includes an article number and a paragraph number like this example from the public health corpus, where the Danish to English translation from Google Translate includes an insertion marked in bold writing:

19a) Reference sentence: 'Den berørte medlemsstat kan give et forskud på mellem 50 % og 100 % af den finansielle støtte, efter at produktions- og afsætningsplanen er godkendt i overensstemmelse med artikel 28, stk. 3, i forordning (EU) nr.'

19b) Candidate sentence: 'The Member State concerned may grant an advance of between 50% and 100% of the financial assistance after the production and marketing plan has been approved in accordance with Article 28 **(2).** 3 of Regulation (EU) No'

This is not an isolated incidence in the translations from Danish to English, in the translations in the public health domain or in the translations from Google Translate as DeepL also makes this error. Another mistake DeepL does is to add an initial '2. ' to the beginning of some sentences in the English to Danish public health translation. These sentences all include a reference to an EU paragraph like the example seen below:

20a) Reference sentence: 'The support referred to in paragraph 1 shall end on 31 December 2020.'

20b) Candidate sentence: '2. Den i stk. 1 omhandlede støtte ophører den 31. december 2020.'

eTranslation is the only tool to not have such issues and the nature of the very few existing examples of insertions are simply a random word inserted on the back of a sentence and not related to any particular style sheet like DeepL and Google Translate. This might indicate a recurring problem for the two tools, whereas it is worth noting that eTranslation has almost no

deletion/insertion errors at all and produce better translations based on this parameter. It would be tempting to claim, that 'of course, eTranslation did not have issues with this', but it would probably be worse for eTranslation not to be able to handle this type of writing style since it is within the realm of the EU.

Miscellaneous:

A category that is less specific is the miscellaneous. No tool had significantly more errors than the others did and the number of errors across the domains were consistently distributed. All three translation systems had wrongly inflected nouns and also translated some sentences very literally as mentioned before. One thing DeepL struggled more with compared to the other two tools are prepositions highlighted in the English to Danish translation of this sentence from the finance domain:

21a) Source sentence: 'In 2009, Rigsrevisionen issued a total of 37 audit opinions on institutions under the Ministry of Economic and Business Affairs, the Ministry of Science, Technology and Innovation, the Ministry of the Environment and the Ministry of Climate and Energy.'

21b) DeepL candidate sentence: 'Rigsrevisionen har i 2009 afgivet i alt 37 revisionspåtegninger på institutioner under Økonomi- og Erhvervsministeriet, Ministeriet for Videnskab, Teknologi og Innovation, Miljøministeriet og Klima- og Energiministeriet.'

Both eTranslation and Google Translate chose the Danish word 'om' instead of 'på', which is a more correct way of translating this particular example. Preposition mistakes happen more than a few times in the translations that DeepL produced and they happen in both languages. It is not something that eTranslation and Google Translate have significantly issues with. I cannot pinpoint exactly where in DeepL's inner workings things are different to the two others regarding these types of errors; also, it is uncharacteristic, since DeepL otherwise displays translations that would lead the reader to assume that these mistakes would not happen.

Entities:

In raw numbers, eTranslation did best regarding entities, which is not a big surprise given that the corpora are from ELRC-SHARE and you would expect an EU tool to perform well on EU institutions, different legislations and so on. The two other tools are also doing well, considering they are not related to some of the entities.

An issue all the tools had was correctly translating the police force of Southern Jutland's name both to English to Danish and the other way around. 'Syd- og Sønderjyllands politi' poses a challenge, since Sydjylland and Sønderjylland are not the same region; in fact, Sydjylland is the area of land that Sønderjylland does not cover in the southern part of Jutland. Therefore, translating the police force's name into 'South and South Jutland Police' or only 'Sydjyllands politi' are not satisfactory translations because of the geographical and cultural implications of the two names. In addition, the prefix 'Sønder-' does not exist in the dictionary in any other form than an expression synonymous with beating something to bits and pieces.

All three tools struggle with identifying when an entity should be translated and when not. An example is the report mentioned in the culture corpus, which is called 'Aarhus European Capital of Culture 2017 – Second Monitoring Meeting' in both the Danish and English references. This was not an issue when translating to English from Danish, however, there were different mistakes when translating from English to Danish. eTranslation was the only tool to consistently translate the name completely to Danish as 'Aarhus Europæisk Kulturhovedstad 2017 – Andet overvågningsmøde' and even though it is an incorrect translation, the consistent translation is a plus. Both Google Translate and DeepL produced the same translation as eTranslation, however, they also produced a hybrid version of Danish and English namely 'Aarhus Europæisk Kulturhovedstad 2017 – Second Monitoring Meeting' showing inconsistent translations, when encountering named entities.

As seen in the error typology scores, entities are one of the main issues. It is not always clear if they have a valid translation in the target language and whether it should be translated or not. A strategy for a tool is to locate/crawl for the organisation's website and see if they have an English/Danish version and pick the name they use themselves. If this is not the case and only one version is available, the systems should stick to that and not attempt to make up a name in the target language.

Languages:

According to the unsmoothed BLEU scores, there is nothing between the two languages as both the English to Danish and Danish to English translations outscore each other on 6 of the 12 translations. The Danish to English translations outscore the English to Danish translations 7 to 4 and Google Translate's general text smoothed translations both score 0,752 resulting in a tie. As mentioned before, there are very few decimal points separating the translations, both language and domain-wise. Both the English and Danish translations are susceptible to Anglicisms and Danish phrasing,

which could indicate an inability to properly understand the full meaning and context of a sentence or not fully understanding the rules of the target language. An example of an Anglicism is from DeepL's English to Danish translation in the public health domain:

22a) Source sentence: 'Availability of the Instrument'
22b) Candidate sentence: 'Tilgængelighed af instrumentet'

Google Translate and eTranslation have both translated this sentence to the more fluent 'Instrumentets tilgængelig'. I have categorised this sort of error as miscellaneous.
Another difference is the way semicolons are handled. A semicolon is a rarity in Danish, due to the strong commas, but they are more common in English. The three tools practised two methods; Google Translate and DeepL simply kept the semicolons from English to Danish, where a comma might have been better and eTranslation simply omitted or replaced them with full stops, which can result in poorly translated sentences.

5 Discussion and perspectives

My aim with this study was to examine the current quality of machine translations from English to Danish and the other way around using three different tools. I highlighted eTranslation on the basis of the opinion of stakeholders asked in the report 'Sprogteknologi i verdensklasse' and this has been compared to the biggest on the market, Google Translate and the most preferred tool by language companies, DeepL. I have compared translations across four different domains to investigate whether domain-specific language would have a significant impact on the quality of machine translation. Furthermore, I have used different evaluation metrics to get a better quality assessment. My findings indicate that the three tools are all outputting translations of very high quality and they could all be useable options for machine translation tasks. In addition, I have demonstrated this for both English and Danish translations across four different domains.

The BLEU scores above 0,6 is a surprise since the narrative in reports has been negative and experts have painted a pessimistic picture. There have not been many recent studies on this topic and reports like the ELRC White Paper call for more research on language technology for Danish. A paper from 2009 has classified Danish to English and vice versa to score >0,5 on a statistical machine translation system. This study also examined pivot language translation and what effect a pivot language could have on translations with low-resource language pairs. An intermediate language could be English or French, which is used as a mediator between two

languages with few parallel corpora and other data available, meaning a Danish-Ukrainian translation is actually Danish to English to Ukrainian. The recent crisis in Ukraine call could call for public sector webpages to be translated into Ukrainian from Danish to accommodate Ukrainian refugees. In such a case, a pivot language would most likely be the best way to translate.

An overview from their study in 2009 can be found in the appendix (Koehn, Birch, and Steinberger 2009). The overview shows that Danish was doing well on statistical machine translation systems, but the transition to neural machine translation systems has lowered the quality and along with the low spending, compared to like-minded languages, has created the disheartened perception. My study contributes to this field by informing stakeholders that Danish machine translations are reliable, although, given the rapid progression and recent initiatives, I would like to see an update in a few years. More and more attention is given to Danish NLP these days and my results contribute to the increasing need for information on what technologies exist and the quality hereof.

The results of the BLEU scores for all the translation systems have surpassed my expectations and the fact that negation errors and made-up words are not problems for neural machine translations anymore is positive. The data follow a trend in 3 out of the 4 domains. The public health domain is the only one with a relatively big variety in BLEU scores compared to the other. As mentioned before, DeepL's scores for both the Danish and English translation of the public health are just above 0,6, which is one of the only points that differentiate the three tools. This trend is visible in the TER scores as well. In theory, the higher the BLEU score is, the lower the TER score should be.

However, the results of the translation edit rates leave something to be desired. They are more varied than the BLEU scores and they indicate that some post-editing is required depending on the domain. The nature of this post-editing can be seen in the errors found This point is backed by the errors located from the error typology analysis, which point to challenges that require knowledge of the world or a specific context in order to produce a correct translation of a term or entity. Something the current machine translation systems do not possess at the time of writing. To iterate a previous point, all the sentences in the candidates are readable and make sense. This is positive since Danish machine translation has reached a threshold and moves the debate to a holistic view of the translation systems. Since the translations are so similar in quality and make roughly the same errors, which factors decide what tool to choose? This might be something to

study in the future as data security, trust, usability, custom vocabularies for commercial systems and maybe even a tool developed by Danes for Danes will see the light of day.

I do not think machine translation systems can substitute human translators, but they are certainly making a strong case with the results. It could be interesting to see to what extent machine translation systems can deliver the same or outperform human translators. The two evaluation metrics account for human correlation by number of references etc. However, the subject of translation is more than similarity and post-editing. Translation is more than a linguistic practice; it also requires knowledge of regions and groups of people in Denmark. A human translator uses translation strategies to transform a piece of text in one language into a different language while retaining meaning, writing style, humour and so on. Machine translations are not human translations, but comparing time, effort with both pre and post-editing, cost and other factors could play into a decision to use a machine translation tool or a human translator. Most translation tasks done by translation companies are done with the help of machine translation, but what is lost and gained? A human translator considers different things and has several micro strategies for a translation. In addition, monoculture specific terms and entities are adapted to suit the receiver. This could be something like a translation of *Middelalderslagsgenfortælling* (middle age battle re-enactment) into *Civil War re-enactment* for an American audience and vice versa. This is one of the major challenges for machine translation and requires more than lexical, morphological and syntactic knowledge (Øveraas 2016).

5.1 Caveats

As mentioned, the lowest score was DeepL's Danish to English public health translation. A possible explanation could be that DeepL did not translate whole documents, but only snippets of around 50.000 characters, as the option to translate documents, is subscription-based. As mentioned, I look at the free versions of the tools. Interestingly, I had the same issue with Google Translate, but upon revisiting Google Translate, I was able to translate whole documents for free, not just parts of them. I had already done some initial analysis on snippets from Google Translate and these were higher than the full document translations. This disproves that length is the reason behind the low score for the public health domain translations of DeepL. The fact that the one system is only snippets of the whole document might invalidate the whole comparison foundation, just as toggling the domain-specific translation function for eTranslation and comparing the output from here instead of the general text domain as I have done.

5.1.1 Evaluation metrics

As mentioned in a previous section, there are drawbacks to BLEU. One is that it only measures similarity and does not account for meaning or syntax as such although n-gram size and smoothing factor can be adjusted. Furthermore, the BLEU score I have used is one-dimensional in its evaluation of whether an n-gram is a match or not and there is not much room for creativity and nuances. In addition, I only used 1 reference to evaluate the quality of the candidate. This means that synonyms and minor differences are penalised. As mentioned previously, there are ways to manipulate the BLEU score using various methods and tricks, which have been discussed by others. In addition, one of the critiques of BLEU is that valid comparisons across all evaluations made with BLEU should be done using the same pre-processing scheme. There are alternatives to BLEU like SacreBLEU and RIBES.

Translation edit rate has limitations as well. According to Snover et al., automatic TER works best with 4 references as it would then correlate on par with human judgment. There is a semi-automatic metric created to better Human-mediated Translation Error Rate (HTER) is a semi-automatic metric that uses references created by humans to acquire better quality estimation. This is done to address some of the limitations of TER, such as dealing with synonyms, something that the metric METEOR is able to as well (Snover et al. 2006). Furthermore, the score generated by TER does not distinguish between the cognitive loads each error has. This means that some errors require more attention to correct than others do and like in the error typology, negations and prepositions are more crucial than some other mistakes. Furthermore, Koponen et al. argue that 38% of post-edits are unnecessary. This argument stems from the fact that human post-editors can have preferences and ‘over-edit’ translations that were good enough already. Also, some errors may not be corrected by the post-editor or new errors may be introduced (Koponen, Salmi, and Nikulin 2019).

Error typology:

An issue with the criteria that a translation should be readable is that it is my assessment of the sentences. What might seem like a readable sentence to one reader might not always correspond to what another reader might deem adequate. For a stronger analysis in the future, professional translators could be employed to carry out the examination. In the excerpts I have analysed for the error typology, the miscellaneous category has been a basket for the ‘rest’. Preposition mistakes are part of this category and they are as disruptive to the meaning as negations and they usually only

make up few words per sentence, therefore, they might not be properly reflected in the evaluation metrics and error typology. The same goes for inconsistent terminology in domain-specific translations. The BLEU and TER scores might be good, but if all the technical terms are translated incorrectly then the rest of the translation is almost invalid.

It was difficult to determine whether an insertion of an article that otherwise would not be present resulted in a meaning misinterpretation or the other way around. Which error influenced the other? Is it the chicken or the egg? It is obvious that insertions or omissions of articles or prepositions result in a misunderstanding, but it is difficult to know where in the algorithm the error has happened. Another way of structuring such an analysis could be to only register if a sentence needed post-editing and thus count correct sentences vs. sentences that need attention.

Corpora:

The translations are only as good as the corpora they are given to translate. This means that even though the corpora were supposed to be parallel corpora, they themselves were translated wrong, which led to translations of the 'same' sentence having different meanings. For instance in the Danish and English culture references, where one person is a racing driver advisor and the other is an advisor on an unspecified topic for racing driver Kevin Magnussen:

23a) English reference: 'What do the Minister of Justice, the car company Jaguar, Prime Minister Lars Løkke and Kevin Magnussen's racing driver advisor have in common?'

23b) Danish reference: 'Hvad har Justitsministeren, bilfirmaet Jaguar, Lars Løkke og racerkører Kevin Magnussens rådgiver tilfælles?'

Differences like this illustrate that the evaluation metrics can function to perfection, but still get it wrong since a perfectly well-translated sentence in this instance would still in theory result in an error due to a meaning misinterpretation. Factors like this can also influence a BLEU score and thus skewing a quality estimation of a translation system. The origin of the corpora also affects the analysis since the corpora may have been translated using machine translation and then post-edited in order to create the parallel corpora. This may have led to Anglicisms in the Danish references and Danish phrasing in the English references.

Another factor that could alter the result of the evaluation is the choice of corpora. The ones I have used are all from the EU and that is the predominant theme in these. Even the general text domain is

quite a formal language compared to newspaper articles or naturally occurring language. I chose the corpora due to their availability and the fact that they are parallel corpora of considerable size. I would speculate that the BLEU score would be lower with a less restrained writing style.

5.2 Future studies

What are the needs for machine translation tasks; Danes are proficient speakers of English, so in what contexts do Danes need translation assistance? According to the ELRC White Paper, there is a demand for translations for EU languages, non-official EU languages and as mentioned before immigrant languages. The paper mentions the Danish public sector as a key stakeholder and highlights municipality websites as concrete cases where translations into other languages both EU languages and Greenlandic and Faroese even though they are not official EU languages. Increased digitalisation, and customisation of domain vocabulary to comply with words like *Fællesoffentlig sector* that do not have a corresponding English term.

Furthermore, valuable Danish data is hard to come by and there is very little coordinated effort to collect language data and spread awareness of how important data and data collection from various sources, like the public sector, can be. It is also important to get Danish sources and not just translated corpora or news articles to train language models (European Language Resource Coordination 2020). As mentioned before, there is a need for Danish NLP enthusiasts and experts. Even at the universities, there is a risk that students of language technology like at the master's programme IT & Cognition put their effort into analysing and utilising the English language. The consequence of this, as mentioned in 'Sprogteknologi i verdensklasse', is that the Danish language with all its challenges both linguistic and language politically is bypassed and thereby further contributing to existing problems for Danish language technology. The high quality of machine translations and other linguistic helping applications are a problem for learners of language in the Danish schools since they do not actually learn a language, but only how to use the tools themselves (Kirchmeier et al. 2019).

Even though we got our hands on more data, it would still need to be marked with metadata in order to be useful for training, processing etc. This requires manual labour and annotation is no cheap task. An initiative created by experts and stakeholders who are trying to progress this is *det Centrale Ordregister (COR)* or The Central Word Register in English. Similar to the Danish social security number, every word will be given a number. This number will be linked to a word and associated information about this word that can be accessed depending on a user's

need, which could be pronunciation, inflexions and so on. This can help track, correct and add new information on Danish words and thus make it easier for everybody involved in Danish NLP. One such area that would benefit is speech recognition, which is difficult in Danish since there is little correlation between spelling and pronunciation. Chatbots that can operate a phone hotline at all times, dictation and virtual assistants are some of the areas that could see huge progress within the next years with a project called KIRI[8]. In addition, how do the virtual assistants deal with anaphoric references and things mentioned previously in the conversation or even ellipsis?

Finally, it could be useful to expand my study and get an overview of how Danish did across all EU languages using eTranslation, Google Translate, DeepL and other big commercial systems like Microsoft Bing Translator and Amazon AWS. This could nuance the results I have found and give a clearer picture of the state of Danish machine translation.

6 Conclusion

In my paper, I have looked at the state of machine translations involving Danish and English using three different translation tools. My point of departure was the report 'Sprogteknologi i verdensklasse' published in April of 2019, where it was stated that one of the areas with the most potential for improvement and applicability was automatic translation in a Danish context.

The results of my analysis show that machine translations involving English and Danish indicate that the Danish language has reached a point where the issues are not a system's lack of training and knowledge of the Danish language, but rather the availability of data and (national) policies on this subject. In fact, the BLEU score evaluations indicated translations of very high quality. The BLEU scores above 0,6 suggest the three systems were able to produce identical translations close to a gold standard/reference. Judging based solely on the BLEU scores suggests that there is no reason to prioritise machine translation in the Danish NLP environment since they perform so well. The translation edit rate examination suggests that there are still some issues, especially with DeepL. Certain domains proved more challenging than others, namely the finance domain and the public health domain when using DeepL.

Overall, the three translation systems, eTranslation, Google Translate and DeepL are all producing high-quality translations in both English to Danish and Danish to English according to the BLEU scores. In terms of usability, they are also very similar, but what truly separates them is the setup

[8] https://videncenter.kl.dk/cases/cases-fra-tekradar-2019/kommune-kiri-faellekommunal-borgerservice-chatbot/

beside the translation system. eTranslation was created by the EU and is now offered to certain stakeholders and promotes themselves on data security among other things. Google Translate is the largest and most known tool and has access to a huge database. DeepL is a commercial tool with services behind a subscription wall and claims to be better than competitors are. I believe this is where the real choice lies for a user of machine translation. However, the intentions and underlying values of the tools are outside the scope of this particular analysis but could serve as a topic within language policy-related studies.

On the matter of domains, I have found domain specificity not to be an issue, not for any of the tools nor the languages concerning the BLEU scores. That being said, the finance domain did present challenges with the technical terms for all the tools. Furthermore, the finance domain consistently scored the lowest or second-lowest BLEU score, the worst or second-worst TER score and was the domain with the most errors in the error typology. The overarching theme of the domains is the EU and EU-formal language, which presented some difficulties for both Google Translate and DeepL, however, eTranslation did well across the board and did not have any issues with the writing style of EU legislations and the like. The domain with the lowest BLEU scores was DeepL's public health domain scores, which is interesting since this domain is the highest scoring domain for the two other translation systems. The biggest differences were revealed in the manual error location analysis.

As previously mentioned, the error types selected for this analysis were issues that had been challenging machine translation systems when translating to and from Danish into English. I can conclude that negation errors are outdated and the only occurrence of a negation mistake was debatable. Likewise, there was only one incidence of a made-up word and that too was questionable, since it was an entity and could have been classified as such. It seems that machine translation has reached a threshold, which is positive. Idioms still prove to be a problem but the nature of neural networks indicates that the translation systems will learn to translate them with time. It is not always the translation system that is at fault for an incorrect translation; they are still human dependent. This is both in regards to the differences in input and accessible data.

The way the neural networks function, learn and solve problems bear resemblance to the human brain, but looking at the results and the issues that are facing them now, I wonder if it is only mimicry and not actual intelligence, since they lack knowledge of the world around them. There is no doubt; however, the transition to neural machine translations has borne fruit and has caught up to

the Danish language. After I saw the results, I was tempted to write the conclusion in Danish and translate it using eTranslation, since I would get the best result from that tool and my data would be safe.

Bibliography

Aiken, Milam. 2019. 'An Updated Evaluation of Google Translate Accuracy'. *Studies in Linguistics and Literature* 3 (3): p253. https://doi.org/10.22158/sll.v3n3p253.

Banerjee, Satanjeev, and Alon Lavie. 2005. 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments'. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics. https://aclanthology.org/W05-0909.

Benjamin, Martin. 2019. 'Empirical Evaluation of Google Translate across 107 Languages'. *Teach You Backwards* (blog). 30 March 2019. https://www.teachyoubackwards.com/empirical-evaluation/.

Breinstrup, Thomas. 2016. 'Vestager tager fat på dataindsamling'. Berlingske.dk. 18 January 2016. https://www.berlingske.dk/content/item/85517.

CEF. n.d. 'ETranslation Builds on MT@EC'. CEF Digital. Accessed 27 October 2021. https://ec.europa.eu/cefdigital/wiki/cefdigital/wiki/display/CEFDIGITAL/eTranslation+builds+on+MT@EC.

'DeepL Press Information | Setting Records!' n.d. Accessed 28 February 2022. https://www.DeepL.com/press.html.

ELIS Research. 2022. 'EUROPEAN LANGUAGE INDUSTRY SURVEY 2022 Trends, Expectations and Concerns of the European Language Industry'.

'ELRC-SHARE'. n.d. Accessed 2 December 2021. https://www.elrc-share.eu/info/#general.

European Commission. 2021. 'Digital Economy and Society Index (DESI) 2021'. https://digital-strategy.ec.europa.eu/en/policies/desi.

European Language Resource Coordination and One Vision Design. 2020. *ELRC White Paper Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe: Why Language Data Matters*.

'Evaluating Models | AutoML Translation Documentation'. n.d. Google Cloud. Accessed 29 November 2021. https://cloud.google.com/translate/automl/docs/evaluate?hl=da.

Kirchmeier, Sabine, Peter Juel Henrichsen, Philip Diderichsen, and Nanna Bøgebjerg Hansen. 2019. 'Sprogteknologi i verdensklasse'. https://dsn.dk/wp-content/uploads/2021/01/sprogteknologi-i-verdensklasse.pdf.

Koehn, Philipp, Alexandra Birch, and Ralf Steinberger. 2009. '462 Machine Translation Systems for Europe'. In *MTSUMMIT*.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. 2019. 'A Product and Process Analysis of Post-Editor Corrections on Neural, Statistical and Rule-Based Machine Translation Output'. *Machine Translation* 33 (June). https://doi.org/10.1007/s10590-019-09228-7.

Larsen, Claus Thornby. 2019. ‘Neural Machine Translation in DA Brief Assessment Report’.

Larsen, Claus Thornby. 2021. ‘ETranslation - Machine Translation at the European Commission’. At University of Copenhagen, Amager.

Martindale, Marianna, and Marine Carpuat. 2018. ‘Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT’. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 13–25. Boston, MA: Association for Machine Translation in the Americas. https://aclanthology.org/W18-1803.

‘MT Quality Metrics’. n.d. Intento. Accessed 16 December 2021. https://help.inten.to/hc/en-us/articles/360020528540-MT-quality-metrics.

‘NLTK :: Nltk.Translate.Bleu_score’. n.d. Accessed 6 December 2021. https://www.nltk.org/_modules/nltk/translate/bleu_score.html.

Øveraas, Kirsten Marie. 2016. *Ti faldgruber: oversættelse for ikke-oversættere*. 1. udgave. Frederiksberg: Samfundslitteratur.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. ‘BLEU: A Method for Automatic Evaluation of Machine Translation’. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, 311. Philadelphia, Pennsylvania: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pedersen, Bolette S., Georg Rehm, and Hans Uszkoreit. 2012. *The Danish Language in the Digital Age*. White Paper Series. Berlin New York: Springer.

Post, Matt. 2018. ‘A Call for Clarity in Reporting BLEU Scores’. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–91. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319.

Sepesy Maučec, Mirjam, and Gregor Donaj. 2020. ‘Machine Translation and the Evaluation of Its Quality’. In *Recent Trends in Computational Intelligence*, edited by Ali Sadollah and Tilendra Shishir Sinha. IntechOpen. https://doi.org/10.5772/intechopen.89063.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. ‘A Study of Translation Edit Rate with Targeted Human Annotation’. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–31. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. ‘Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric’. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 259–68. Athens, Greece: Association for Computational Linguistics. https://aclanthology.org/W09-0441.

Turovsky, Barak. 2016. ‘Found in Translation: More Accurate, Fluent Sentences in Google Translate’. Google. 15 November 2016. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

‘Understanding MT Quality: BLEU Scores’. n.d. Accessed 13 September 2021. https://www.rws.com/blog/understanding-mt-quality-bleu-scores/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. ‘Attention Is All You Need’. *ArXiv:1706.03762 [Cs]*, December. http://arxiv.org/abs/1706.03762.

‘What Is ETranslation’. n.d. CEF Digital. Accessed 1 December 2021. https://ec.europa.eu/cefdigital/wiki/cefdigital/wiki/display/CEFDIGITAL/What+is+eTranslation.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. ‘Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation’. *ArXiv:1609.08144 [Cs]*, October. http://arxiv.org/abs/1609.08144.

Appendix

	Target Language																				
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	57.2	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	41.0	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	42.7	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Table 3: Translation performance as measured in %BLEU for all 462 language pairs

Figure 3 Statistical machine translation performance from (Koehn, Birch, and Steinberger 2009)

EN-DA			**DA-EN**		
eTranslation	Google Translate	DeepL	eTranslation	Google Translate	DeepL
Public Health					
Without smoothing					
0,818	0,780	0,621	**0,821**	0,788	0,615
With smoothing					
0,909	0,876	0,759	**0,925**	0,909	0,751
Culture					
Without smoothing					
0.728	0,757	**0,770**	0,743	0,759	**0,774**
With smoothing					
0,758	0,784	**0,906**	0,772	0,786	**0,900**
General text					
Without smoothing					
0,706	**0,722**	0,719	0,707	**0,718**	0,690
With smoothing					
0,737	0,752	**0,834**	0,741	0,752	**0,806**
Finance					
Without smoothing					
0,685	**0,709**	0,708	0,679	0,684	**0,692**
With smoothing					
0,775	0,798	**0,823**	0,828	**0,833**	0,795

Table 17 Comparison of unsmoothed and smoothed BLEU scores across the tools and languages

Error types	Inconsistent terminology	Omissions/ additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	39	22	4	0	167	23
Culture (82)	0	4	0	0	20	3
Finance (164)	35	7	4	0	56	20
General (23)	6	2	0	0	23	2

Table 18 Domain specific translation error types for English to Danish translations from eTranslation

Error types	Inconsistent terminology	Omissions / additions	Made-up words	Negation error	Miscellaneous	Named entities
Public Health (266)	56	8	3	0	112	44
Culture (82)	0	8	0	0	24	0
Finance (164)	28	12	2	0	68	12
General (45)	7	1	0	0	12	0

Table 19 Domain specific translation error types for Danish to English translation from eTranslation

Google Trans General Text English Candidate

5. Notification of near misses and damage to facilities as well as evacuations

The operator or owner is obliged to report incidents and damage to the plant as well as evacuations to the Danish Working Environment Authority.

In cases where notification must be made immediately by the South and Southern Jutland Police or by the Danish Working Environment Authority pursuant to section 8 of the Executive Order on Notifications, it is the operator or the owner who has the duty to notify.

Overview of notification of near misses and damage to the facility as well as evacuations

The table provides an overview of which near misses and evacuations must be reported, when and to whom.

The table includes notifications to the South and Southern Jutland Police and the Danish Working Environment Authority under the headings of notification.

Notes to the table:

a. describe the course of events and describe what preventive measures have been taken so that recurrences are avoided.

From 19 July 2018, the notification with associated information must be made in accordance with the common reporting format set out in Annex I, incident types AF, and I, to Commission Implementing Regulation (EU) No. 1112/2014 of 13 October 2014, cf. Appendix 3 in the Notice of Notice.

From 19 July 2018, these operators and owners must report the incidents with associated information in accordance with the common reporting format set out in Annex I, incident type AF, and I, to the Commission's Implementing Regulation (EU) No. 1112/2014 of 13 October 2014, cf. Appendix 3 in the Notice of Notice.

To Commission Implementing Regulation No 1112/2014, a guide has been prepared, which aims to provide the competent authorities, operators and owners with additional information and examples in order to promote a uniform interpretation of the reporting requirements in the Implementing Regulation.

See more under the section "Background" in this guide.

It appears from the executive order on medical control of work with ionizing radiation in connection with offshore oil and gas activities, etc., when it must be reported that a person has been exposed to ionizing radiation.

Notification of incidents that have or could have resulted in the release of biological agents is stated in the Executive Order on protection against exposure to biological agents in connection with offshore oil and gas activities, etc.

It is the responsibility of the operator or owner, respectively, to make a competent assessment of the individual near -incident with a view to whether it should be reported.

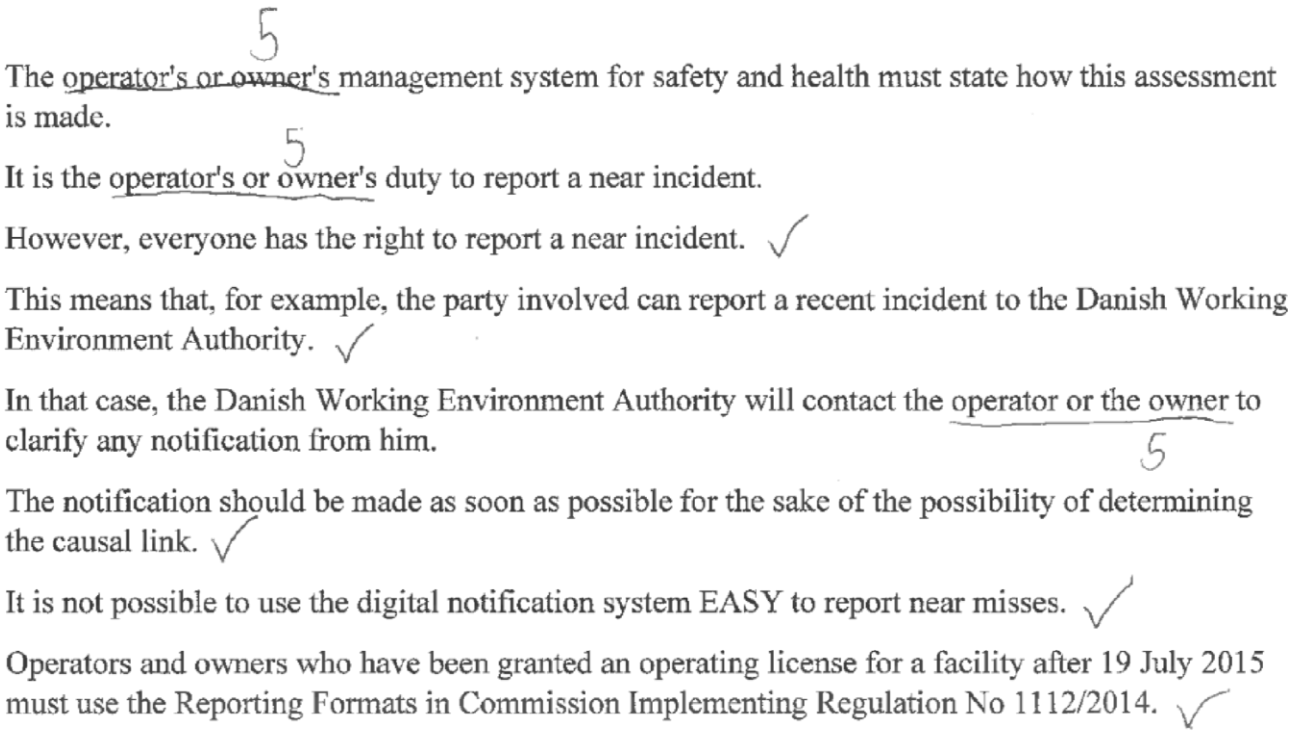

Figure 4 Example of error categorisation for the error typology