Comparing Encoder-Decoder Architectures for Neural Machine Translation: A Challenge Set Approach

Coraline Doan

A thesis submitted in partial fulfillment of the requirements for the
Master's degree in Translation Studies

Under the supervision of
Dr. Elizabeth Marshman

School of Translation and Interpretation
Faculty of Arts
University of Ottawa

Abstract

Machine translation (MT) as a field of research has known significant advances in recent years, with the increased interest for neural machine translation (NMT). By combining deep learning with translation, researchers have been able to deliver systems that perform better than most, if not all, of their predecessors. While the general consensus regarding NMT is that it renders higher-quality translations that are overall more idiomatic, researchers recognize that NMT systems still struggle to deal with certain classic difficulties, and that their performance may vary depending on their architecture. In this project, we implement a challenge-set based approach to the evaluation of examples of three main NMT architectures: convolutional neural network-based systems (CNN), recurrent neural network-based (RNN) systems, and attention-based systems, trained on the same data set for English to French translation. The challenge set focuses on a selection of lexical and syntactic difficulties (e.g., ambiguities) drawn from literature on human translation, machine translation, and writing for translation, and also includes variations in sentence lengths and structures that are recognized as sources of difficulties even for NMT systems. This set allows us to evaluate performance in multiple areas of difficulty for the systems overall, as well as to evaluate any differences between architectures' performance. Through our challenge set, we found that our CNN-based system tends to reword sentences, sometimes shifting their meaning, while our RNN-based system seems to perform better when provided with a larger context, and our attention-based system seems to struggle the longer a sentence becomes.

Keywords: neural machine translation, machine translation evaluation, convolutional neural network, recurrent neural network, attention-based neural machine translation, challenge set

Résumé

La traduction automatique (TA) comme champ de recherche a connu des avancées considérables au cours des dernières années grâce à l'intérêt croissant pour la traduction automatique neuronale. En combinant l'apprentissage profond à la traduction, les chercheurs ont pu livrer des systèmes offrant une performance supérieure à la plupart de leurs prédécesseurs, voire tous ceux-ci. Bien que le consensus général concernant la TA neuronale soit qu'elle produit des traductions de qualité supérieure qui, dans l'ensemble, sont plus idiomatiques, les chercheurs reconnaissent que les systèmes de TA neuronale peinent encore à traiter certaines difficultés classiques et que leur performance pourrait varier en fonction de leur architecture. Dans le présent projet, nous adoptons une approche dite de « challenge set », soit formée d'exemples de phrases illustrant une série de difficultés, pour évaluer des exemples des trois principales architectures de TA neuronale, soit les systèmes fondés sur les réseaux de neurones à convolution, les systèmes fondés sur des réseaux de neurones récurrents (RNR) et les systèmes basés sur l'attention, tous trois entraînés sur le même ensemble de données destinées à la traduction de l'anglais vers le français. Le « challenge set » comporte une sélection de difficultés d'ordre lexical et syntaxique (p. ex. des ambiguïtés) tirées de références sur la traduction humaine, sur la TA, ainsi que sur la rédaction à des fins de traduction, et comprend aussi des variations en longueur de phrases et en structures qui sont reconnues comme étant des sources de difficultés, même pour des systèmes de TA neuronale. Ce « challenge set » nous permet d'évaluer la performance en général de tous les systèmes en fonction de diverses difficultés et nous permet également d'évaluer leur performance en fonction de leurs différentes architectures. Grâce à ce « challenge set », nous avons constaté que notre système fondé sur les réseaux de neurones à convolution a tendance à reformuler les phrases, changeant parfois leur sens, alors qu'un contexte plus long semble favoriser notre système fondé sur les RNR, tandis que notre système basé sur l'attention semble trouver les phrases de plus en plus difficiles à traduire plus elles sont longues.

Mots-clés : traduction automatique neuronale, évaluation de la traduction automatique, réseau de neurones à convolution, réseau de neurones récurrents, traduction automatique basée sur l'attention, *challenge set*

Acknowledgements

It goes without saying that this research project wouldn't have been possible without the help and support that was so graciously offered by the National Research Council of Canada (NRC). I would like to thank the Digital Technologies Team and especially Roland Kuhn for introducing me to the NRC, Samuel Larkin, for teaching me all things NMT, and Cyril Goutte for providing me with his expert advice on statistics. Thank you for your patience and for always being willing to answer the numerous questions I had.

I would also like to thank my supervisor, Dr. Elizabeth Marshman, for her immense support and guidance throughout this entire process. I cannot thank you enough for all the work you've put in to ensure that I would be proud of my thesis. You've supported me through the ups and downs of carrying out this research, and always found ways to help me overcome any obstacle I encountered. You've shown incredible compassion while maintaining your professionalism and I will forever be grateful to you.

Thank you to my past colleagues from the Canada Revenue Agency and to my new team at the Public Health Agency of Canada for their support and their encouragement. Thank you for being so flexible and for watching me grow as a professional throughout this degree (and pandemic!). I would like to thank three colleagues in particular: Jason Lawson for believing in me, Sheriff Abdou for giving me a chance, and my current director, Susan Ternan for continuously supporting me and for always being so understanding.

Thank you to the friends I made in Sherbrooke for becoming my second family, to my new friends from uOttawa for always keeping me motivated, and to the friends I made before starting my university journey for always supporting me no matter the distance.

Lastly, thank you to my family and to my boyfriend for being there and supporting me, some financially, some morally, but all with immeasurable love.

Table of Contents

List of tables

List of figures

List of abbreviations

ALPAC	Automatic Language Processing Advisory Committee
AWS	Amazon Web Services
BLEU	bilingual evaluation understudy
CAT	computer-aided translation
CNN	convolutional neural network
FAIR	Facebook AI Research
GF	gap filling
GPU	graphics processing unit
GRU	gated recurrent unit
LDOCE	*Longman Dictionary of Contemporary English*
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MT	machine translation
NIST	National Institute of Standards and Technology
NLP	natural language processing
NMT	neural machine translation
NRC	National Research Council of Canada
PE	postediting
POS	part of speech
RBMT	rule-based machine translation
RNN	recurrent neural network
SL	source language
SMT	statistical machine translation
LSTM	long short-term memory
TAUM	Traduction Automatique à l'Université de Montréal
TB	Translation Bureau
TL	target language
TM	translation memory
TMS	terminology management system

Introduction

This research was carried out with the goal of gaining insight into different encoder-decoder neural machine translation (NMT) models from a translator's perspective. As such, it focuses first and foremost on the subtleties of language that researchers from other fields may overlook when developing NMT systems. Its findings may help these researchers to improve their models, and also help sensitize language professionals to the kinds of results and challenges they may expect when using NMT.

Context and motivation

Research and development in machine translation (MT) has grown considerably with the popularization of NMT in 2016. Although this new approach seemed to have enhanced the quality of MT, these systems are still not capable of translating flawlessly. NMT helps reduce the administrative burden of maintaining and updating the systems that other approaches relying on rules or statistics entail, but its use or implementation is nonetheless questioned by language professionals and employers alike. The main issue with NMT is that when a translation error occurs, it is almost impossible to track or pinpoint the source of that error. Since the system is self-learning and acts as a "black box", we have less control and oversight over what it assimilates (Kenny, 2018, p. 438). While it is harder to fix a specific or recurrent issue in NMT systems, world-leading companies and institutions continue to invest into the technology because NMT systems most often produce translations that are more accurate and fluent than its predecessor, statistical machine translation (SMT) (Wu *et al*., 2016, p. 20).

As of the writing of this thesis, there are three architectures considered to be state-of-the-art in their field: attentional recurrent neural networks, self-attentional transformers, and fully convolutional networks (Hieber *et al*., 2017, p. 1). Current research in NMT therefore focuses on developing and testing different approaches to each of these architectures. Over the years, many evaluation methods have been proposed to measure the different systems' overall performance, to compare them, and ultimately to attempt to improve them.

In our research, we will be using an approach inspired by Isabelle, Cherry, & Foster's (2017) challenge set evaluation method on three experimental NMT models representing the three dominant architectures, using as a reference and point of comparison two large-scale commercial NMT systems and a hybrid phrase-based statistical machine translation (PBSMT)[1] system (an example of the most popular SMT architecture). We will try to establish a correlation between how these systems were built and how they perform when they are faced with specific, widely recognized lexical or syntactic challenges. By having our systems tackle isolated challenges, we are hoping to identify trends and/or patterns from a translator's perspective that could perhaps lead to improving these NMT systems, and by presenting the strengths and weaknesses of each of the systems, we are hoping we can help MT users make an informed decision when it comes to using MT in a professional environment, whether it be for gisting (i.e., getting a general idea of the meaning expressed) or postediting (PE) purposes. Furthermore, by evaluating MT's handling of recurrent phenomena, rather than of specific items, we are hoping to create a challenge set that can be recycled and reused to evaluate future models or new emerging architectures.

The idea of introducing MT to the workplace often seems to spark a debate between two opposing "teams": those against the use of MT because of the lack of quality assurance, and those who preach it over human translation because of the speed and reduced costs. Our goal is to identify the different contexts in which a certain system architecture may be more appropriate than another, based on realistic insights into MT's potential and shortcomings. We want to present MT as a tool, as an assistant, rather than a solution or a replacement, and to help to explore the potential impact it could have.

Brief overview of MT

Throughout the years, there have been many applications of MT, the three main paradigms being rule-based machine translation (RBMT), SMT, and NMT.[2] Back in the 1950s, the first functional MT system could translate sentences from Russian to English

[1] Also referred to as "phrase-based machine translation (PBMT)"

[2] Another approach, example-based machine translation (EBMT) (e.g., Hutchins 2005), while it produced interesting results, never became a dominant paradigm in MT.

with the help of a dictionary and some linguistic rules (Hutchins, 2004, p. 1). As its name suggests, RBMT relies on rules (i.e., formal representations of word forms and sentence structures and how they may be manipulated) established and manually input by humans. The process is complicated because developers have to take into account the various grammatical, morphological, syntactic, and lexical rules for each language that they wish to translate from or to. Although this first RBMT system was a breakthrough, it also created unrealistic expectations for MT development, as the general public was under the impression that achieving MT of good quality was closer than it really was. This belief had both a positive and a negative impact: on the one hand, research in MT gained more interest and received more funding as a result; on the other hand, people were looking forward to a technology that could not have possibly produced results meeting their expectations with the hardware available at that time. This led to the creation of the Automatic Language Processing Advisory Committee (ALPAC) in the United States, which was put in charge of evaluating the prospects of MT.

In 1966, they released a report that stated that "MT was slower, less accurate and twice as expensive as human translation" (Hutchins, 2001, p. 6). This essentially halted research in MT in the United States, although countries in the European Union and Canada still had a strong need for automatic translation. This need became particularly pressing in Canada when the *Official Languages Act* was passed in 1969, requiring all federal institutions to send out communications to the public in both official languages (English and French), simultaneously. To help fulfill this requirement, a team of researchers from the Université de Montréal (Traduction Automatique à l'Université de Montréal, TAUM) developed the first fully functional RBMT system to translate weather reports. The successful launch of their system inspired others to look into the RBMT architecture, leading to the development of three main approaches: the direct approach (used by the team from Montréal), the interlingua approach, and the transfer approach.

Hutchins (2001) recounts, however, that as the years went by, researchers found that systems relying on an RBMT architecture were very hard to build: they were too ambitious, needed to fulfill too many requirements, and had to be tweaked too often. Consequently, they started to look into SMT, as statistical methods had previously proven to be successful in speech recognition. Unlike RBMT, the approach of SMT relies not on

producing one exact translation, but rather on studying large collections of previously translated texts to identify possible equivalents and word sequences based on these examples, generating thousands of possible translations, and then ranking them based on their predicted degree of correctness. This meant that a large quantity of human translations was required, but that translations rendered by SMT systems were based on these human translations and, therefore, could sound more idiomatic than translations from RBMT systems. It should be noted, however, that, unlike RBMT systems, SMT systems do not have any explicit knowledge of grammar, which can sometimes result in inconsistencies between translations.

Several approaches were tested for SMT, notably word-based models and phrase-based models. As their name suggests, word-based models translated sentences word by word, while phrase-based models translated chunks of of the sentences at a time. Overall, SMT performed much better than RBMT if given sufficient training data. Developed in a subsequent step, hybrid SMT systems (i.e., SMT systems enhanced by the addition of grammar rules that the original systems were lacking), became the most used architecture. However, SMT systems were not only complicated to build and maintain (requiring a lot of human assistance), but also often had to make use of a pivot language for translation between language pairs where bilingual data was insufficient, and this ultimately led to a move towards NMT (Geitgey, 2016).

Computer scientists began looking into a self-learning model for MT and the idea became reality in 2014, when a team of researchers combined recurrent neural networks (RNNs) and encodings (Geitgey, 2016). With RNNs, the system can learn from every input it receives and with an algorithm, the system can figure out the grammatical, morphological, syntactic, and lexical rules that, in previous models such as RBMT and SMT, needed to be manually established and added, or were simply absent from the systems. As a result, much less maintenance is required and the system rarely needs to be manually updated. Furthermore, NMT systems seem to receive higher Bilingual Evaluation Understudy (BLEU)[3] scores when compared to RBMT and SMT systems (Wu *et al.*, 2016, p. 3), which encouraged researchers to follow the neural trend. Thus,

[3] More details on this evaluation method are provided in 1.2.1

research in NMT grew exponentially in the past years and is expected to grow even more as new approaches are tested for improved quality, speed, and idiomaticity.

Current context of MT use

While there has been significant improvement in the field of MT in the last decades, the use of MT in a professional context is still being questioned, in part because of lack of quality assurance. However, although MT is rarely used without some kind of human intervention for publication purposes, it can still be used as a tool for gisting or PE. For example, the Translation Bureau (TB), in collaboration with the National Research Council Canada (NRC), first implemented an SMT system on the Government of Canada's intranet in 2016, indicating that it was to be used for gisting and to help people understand their second official language. The tool, referred to as the "Translation Bureau's Language Comprehension Tool," is an earlier version of Portage (more precisely, version 4.0.3 as of the writing of this thesis) and can translate text of up to 2,000 characters (about 500 words). It is in no way meant to produce translations of publishable quality; in fact, the TB "recommends using this tool for the purposes of improving the understanding of short, simple and unofficial communications in [one's] second language. (…) This tool should not be used by public servants for official publications and outgoing correspondence" (Public Works and Government Services Canada).

Indeed, MT for gisting is not meant to produce high-quality translations; the system is expected only to be accurate enough that the ideas expressed in a source sentence or text are translated in a way that is intelligible to a speaker of the target language. In these cases, unassisted MT can be useful (García, 2009, p. 206), as style is not very important.

In situations where high-quality translation is required or desired, PE of MT (sometimes referred to as PEMT) can be used. PE, as the name suggests, is the process of editing a text after it has been translated entirely by a machine. Much of the research done on PE focuses on productivity (Koponen & Salmi, 2015, p. 118). Researchers compare the process of translating a text from scratch to the process of reviewing a machine-translated text, mainly looking at ways to reduce the amount of time and effort

required (or perceived) for translation (Gaspari *et al.*, 2014; Macken *et al.*, 2020). However, unlike for gisting, productively implementing PE of MT requires an MT system that already produces high-quality translations, which means that unassisted, generic (including free online) MT may not be recommended for PE and in-house products are often required. Indeed, having to revise a poorly machine translated text could in fact add more time to the overall translation process, as it can require more cognitive effort on the translator's part (Sun, 2019, pp. 148-149).

Nevertheless, interest in using MT in a professional context has without a doubt increased with the development of new, promising models and thorough evaluations of each model are necessary, regardless of their intended use.

Our focus

The number of approaches to NMT is rising, but as previously mentioned, three models remain at the top.[4] RNN-based systems, CNN-based systems, and attention-based systems. While we know the strengths and weaknesses of each model in terms of their training time and hardware requirements, we do not have an exact idea about the degree to which there are differences in the quality of their outputs that are noticeable for human users. Therefore, by examining output of examples of these types of systems, we are hoping to see if a system's architecture will noticeably influence its performance. While there are many studies comparing the different NMT architectures (Lakew, 2018; Domhan, 2018), these studies mostly do so by using automatic metrics such as BLEU scores only. While BLEU, which automatically compares MT output to one or more reference human translations, is a quick, inexpensive, and language-independent way to rate MTs (Papineni *et al.*, 2002, p. 311), using BLEU metrics alone to evaluate the quality of a translation is insufficient.[5] In their paper titled "Re-evaluating the role of BLEU in Machine Translation Research," Callison-Burch, Osborne and Koehn noted that "there are millions of variations on a hypothesis translation that receive the same Bleu [sic] score" and that "there are translations which have the *same* Bleu score but *worse*

[4] As of the start of this project

[5] BLEU scores will be further detailed in Section 1.2.1

human evaluation" (Callison-Burch *et al*., 2006, p. 1). These realizations reinforce the fact that human evaluation of MT is still very valuable and needed.

When it comes to evaluating the quality of a translation, Hutchins and Somers believe there are three main elements to take into account: accuracy, clarity, and style (Hutchins & Somers, 1992, p. 163). In MT research where human evaluators are involved, sentences are often given a score based on all of these elements. While this approach is useful to rate a system's overall performance and compare it to another, it provides little information on a system's specific strengths and weaknesses, and gives insight into how to improve a system.

An approach to MT evaluation that aims to fill these gaps is the challenge set, an approach that involves testing systems on a set of sentences specifically designed to illustrate well-known (machine) translation challenges. Using a challenge set helps us target specific error types and to begin to predict how likely they are to be successfully handled by a system. Using multiple examples of the challenges helps to determine whether a mistranslation is a one-time occurrence or a recurrent phenomenon. Consequently, this will tell us the areas on which we ought to focus, should we want to tweak a system. This error-focused approach is especially interesting to us because of our translation background.

Over the years, we have noticed recurring lexical and syntactic structures that have proven to be problematic in an English to French translation context. Some of these difficulties have been addressed, for example, by Jean Delisle and Marco Fiola in *La traduction raisonnée*, 3rd edition,[6] the latest version of the text considered by many to be the reference manual *par excellence* for teaching English to French translation in Canadian institutions, and also a useful guide for professional translators.[7] Renaud-Bray, the largest chain of French-language bookstores in North America, for example, describes *La traduction raisonnée* as "Un classique" and "L'ouvrage indispensable en traduction" on its website (Renaud-Bray, n.d.). The text categorizes difficulties into

[6] Delisle was the author of the first two editions of *La traduction raisonnée* (1993, 2003), with the participation of Alain René in the second edition.

[7] In André Sénécal's (n.d.) words, "Enfin, les traducteurs d'expérience devraient garder *La traduction raisonnée* à portée de la main comme une référence crédible et précieuse, ne serait-ce que pour y trouver l'étincelle d'inspiration susceptible de les relancer dans les moments difficiles de leur pratique."

objectives, presents examples, and provides solutions for each of the problems. Those familiar with the learning method can easily identify very specific turns of phrase in translated texts that suggest a translator has been trained with *La traduction raisonnée*. This shows just how widespread the book's objectives are, and this is also the reason why we used some of the difficulties identified by the authors as the initial inspiration for our challenge set. Our reasoning is that, if a human can detect a recurring translation difficulty and apply a systematic wording to solve it, a machine translation system learning from human translations should be able to do the same.

To complement the book's objectives, which target translation difficulties humans might find challenging, we also looked into translation difficulties that are known to be difficult for machines to tackle. For this reason, our reference sources include works not only from translation pedagogy (namely *La traduction raisonnée*), but also from writing for translation, translatability, and computers and language.

Hypotheses

Based on our current knowledge of the different NMT architectures and on other studies that have touched upon similar subjects, we are proposing two research hypotheses. The first is that different NMT system architectures will perform differently when faced with certain types of challenges. This hypothesis is based on the fact that very different approaches were adopted when envisioning and developing these architectures. RNNs, for example, have been found to work very well with linear structures and to be better at syntax-related tasks (Yin *et al.*, 2017). CNNs, on the other hand, may be better at lexically related tasks (Yin *et al.*, 2017), as they are non-linear and may be better at capturing a sentence's context, while attention might allow for better translations of longer sentences (Tang *et al*., 2018).

The second hypothesis is that, although CNNs have their advantages, the CNN-based system will most likely perform more poorly than the RNN or attention-based systems. This comes from the fact that CNN-based systems are not known to be strong in terms of achieving accuracy in long-range dependencies, i.e., when the cues that a system must recognize to produce a correct analysis and translation of a sentence element are found some distance away, as is the case of many of our challenges, and we are first and

foremost looking at accuracy in our sentences (Tang *et al.*, 2018). This being said, while we are anticipating differences between the systems, we will not be able to fully determine if there is one system better than the others, as this is only an exploratory approach to studying translation difficulties and their handling in the different systems.

Basic methodology

We started by creating a new challenge set for English to French translation based initially on objectives described in Jean Delisle and Marco Fiola's *La traduction raisonnée*, 3rd edition, and expanded to integrate phenomena widely recognized as being difficult for (machine) translation. We identified two lexical difficulties and three syntactic difficulties that we deemed could reasonably be tackled by a machine translation system (including ambiguities, anaphora, and scope of conjunction and modification), and formulated eight sentences for each of the difficulties (or sub-categories of the difficulty) so as to clearly reflect the translation problem. We then manually translated those sentences (144 in total) as reference translations.

Meanwhile, with the help of the NRC, we obtained access to three NMT systems (RNN-based, CNN-based, and attention-based), as well as a hybrid SMT system with a neural component (Portage). All four systems were trained on data from the WMT18,[8] more specifically with four general-language corpora. Training stopped once all four of our systems achieved comparable BLEU scores. That is when we entered our challenge set into the systems and assessed how each of them performed by either marking the translations as correct (green checkmark ✓) or incorrect (red ✗mark) (see section 3.2 for more details).

Subsequently, we analyzed each difficulty and tried to identify patterns or trends in the results, with the goal of establishing correlations between the known strengths and weaknesses of the various architectures and their overall performance for each type of difficulty.

[8] WMT used to refer to the Workshop on Statistical Machine Translation, which has now been renamed Conference on Machine Translation. The already-established acronym WMT was kept to refer to the Conference, where datasets are provided and MT-related tasks are listed for researchers to take on, in a competitive spirit.

Structure of the thesis

The remainder of this thesis will be divided into four chapters followed by a conclusion. In Chapter 1, we will present the evolution of MT by going through the main milestones: from the Georgetown-IBM experiment, to RBMT, computer-aided translation (CAT) tools, SMT, and finally, NMT. We will then discuss some well-established NMT architectures, i.e., RNN-based, CNN-based, and attention-based, as well as give a quick overview of hybrid SMT models. We will close Chapter 1 by going over evaluation of MT, both automatic and human, and we will introduce challenge sets to assess NMT.

In Chapter 2, we will describe our methodology. We will give a description of Portage and discuss the steps required to develop our NMT systems, including selecting our training data and preparing it for system training, which includes normalization, tokenization, and word embedding. We will then provide our systems' specifications.

In Chapter 3, we will present the design of our challenge set, discuss the choices we made in selecting the items included, and explain our evaluation method. We will present all 18 items and provide reference translations for the reader to get an idea of the types of translations we anticipated.

Chapter 4 will be our results and analysis. We will go over all the sentences from all the items in our challenge set, discuss the systems' performance, elements we were expecting and others that took us by surprise. The first two sections will be dedicated to the results for our lexical and syntactic difficulties, respectively, and the third will be a summary of our key findings, where we will also try to identify patterns that seem to be system-specific.

Finally, in our conclusion, we will go through our research questions and hypotheses and address them. We will discuss the limitations and advantages of adopting a challenge set approach and we will suggest future work, including improvements to our own challenge set.

Chapter 1: Literature review

In this chapter, we will be discussing the main paradigms of machine translation (MT) and computer-aided tools (CAT) tools, how they are perceived, and how they are used. We will be presenting this information in chronological order to illustrate the different shifts in approaches, as well as to provide explanation as to why research in MT is where it is now. More specifically, we will be discussing three models of MT and their main sub-types.

1.1. Evolution of machine translation

The field of MT and its history have been thoroughly and ably described by Hutchins (1995, 2001, 2004) and Hutchins & Somers (1992), from which we have drawn a number of key milestones and concepts for this section. As they describe, the idea of MT was first system capable of translating sentences from Russian to English on an IBM 701 computer (Hutchins, 2004, p. 3). Their system relied on a 250-word dictionary, had six grammar rules, and mainly translated sentences in the field of chemistry. Although the system was not as much of a prototype as it was a showcase (p. 10), the results presented were nonetheless impressive and drew the attention of the press: automatic translation appeared to be feasible for the first time. Soon after that experiment, MT became a recognized field of research and, as the number of MT researchers increased, financial support from various organizations grew as well.

1.1.1. Georgetown-IBM experiment: The aftermath

As Hutchins (2004) recounts in his history of MT, while the Georgetown-IBM experiment showed that experimental MT was possible, it also gave unrealistic expectations for MT development. The general public was under the impression that achieving high-quality MT was closer than it really was. The results in the decade following the Georgetown-IBM experiment were disappointing, which prompted the United States government to set up the Automatic Language Processing Advisory Committee (ALPAC) to examine the prospects of MT. In 1966, the infamous ALPAC report was published and research in MT hit a wall: MT was deemed slower, less

accurate, but more expensive than human translation (Hutchins, 2001, p. 6). Instead, it was suggested that researchers focus on developing machine aids for translators, or CAT tools.

Inevitably, research in MT became less and less popular in the United States following the publication of the report. However, the demand in scientific and technical translation remained high in European countries and in countries such as Canada (Hutchins, 2001, p. 7), where bilingualism/multilingualism and linguistic duality/plurality are very present, and some research into MT continued.

1.1.2. Rule-based machine translation

This early research into MT focused mainly on different strategies for rule-based machine translation (RBMT). RBMT, as its name suggests, relies on rules and dictionaries. To develop an RBMT system, developers have to establish various grammatical, morphological, syntactic, and lexical rules (i.e., explicitly describe how words and sentence structures can be formally represented) that define each language that they wish to translate to or from, in addition to rules governing the manipulations required to transform an SL structure into a TL structure (e.g., transfer rules). Combined with bilingual dictionaries, those rules are meant to—more or less—imitate a human linguist's thought process when analyzing a source sentence, translating it, and producing a target sentence.

The first RBMT system to demonstrate promising results was TAUM-MÉTÉO. The project started because, in 1969, Canada's *Official Languages Act* came fully into force, requiring federal institutions to send out communications to the public in both official languages, simultaneously. Quickly enough, Environment Canada found that they were facing an issue when they needed to issue weather warnings: in urgent situations such as when a tornado is approaching, they need to be able to send out many updates in a short period of time, which, consequently, greatly increases the number of high-priority translation requests (Gotti *et al.*, 2014, p. 404). To address this issue, a group of researchers at the Université de Montréal (Traduction Automatique à l'Université de Montréal, or TAUM) developed a prototype MT system capable of translating weather forecast bulletins and alerts from English to French. The system, named TAUM-MÉTÉO,

was launched in the mid-1970s, relied on three bilingual dictionaries and—thanks to its highly restricted vocabulary and syntactic rules—was able to accurately translate weather reports.

Research in RBMT remained popular throughout the 1970s and continued until the end of the 1980s. During those years, there have been three predominant designs for RBMT systems: the direct translation approach (on which TAUM-MÉTÉO relied), the transfer approach, and the interlingua approach.

1.1.2.1. Direct approach

Direct translation was the most rudimentary approach. It relied on dictionaries and basic rules, and the systems produced using this approach were unidirectional and designed for a specific pair of languages (i.e., from a specific source language [SL] to a specific target language [TL]) only. Considered as the "first generation" of MT (Hutchins, 1995, p. 3), systems using a direct translation approach aimed to translate with "minimal amount of analysis and syntactic reorganisation" (Hutchins, 2001, p. 3). Given this approach, translation solutions that are as similar in structure and equivalence were vastly preferable to other options. Words in the SL were reduced to their most basic, uninflected forms (what would today be equivalent to lemmatization in morphological analysis) and were then looked up in a bilingual dictionary, where a TL equivalent would be given. This worked in some cases, as seen with the TAUM-MÉTÉO system. However, achieving good overall performance required not only that the pair of languages have a similar syntactic structure, but also that the field in which the researchers were translating be very specific. Indeed, this type of RBMT system can only be achieved if the vocabulary is limited to a specific domain, for example meteorology in the case of TAUM-MÉTÉO. In fact, after the successful launch of TAUM-MÉTÉO, the TAUM team attempted to recreate the experiment in the field of aviation, but the project was terminated due to the prevalence of complex noun compounds and phrases found in aviation terminology that would have been difficult to translate even for a human translator (Hutchins, 1995, p. 10). Since very strict requirements had to be met in order for the direct translation approach to work, the method was quickly dropped and researchers focused on a more abstract design: the interlingua approach.

1.1.2.2. Interlingua approach

The interlingua approach is based on the idea that there could be an interlingual language, consisting of codes or symbols, independent of both the SL and the TL, but capable of linking the two. As opposed to the direct approach, the interlingua approach breaks the translation process into two steps: from the SL to the interlingual language, then from the interlingual language to the TL (Hutchins, 2001, p. 3). One could also view the interlingual language as an abstract representation of both the SL and the TL. As mentioned earlier, the interlingual language links the SL and the TL. However, considering that the approach is unidirectional, the interlingual language can be seen as the "projection" of the SL, while also being the "basis for the generation" of the TL (Hutchins & Somers, 1992, p. 73). This facilitates the development of multilingual systems, since the addition of a new language only requires two new modules: an analysis grammar (i.e., a set of rules that allows for sentences in the new SL to be converted to its interlingual representation) and a generation grammar (i.e., a set of rules that allows for the interlingual representation to be converted to sentences in the new TL) (Hutchins & Somers, 1992, p. 74), both only specific to the additional language.[9] Researchers were able to add language pairs exponentially with this method, e.g., a bilingual system would have two language pairs, but a trilingual system would have six, and so on.

Although this approach seems to have many advantages, it is, by nature, a very ambitious one. To have a universal interlingual representation that is capable of connecting languages that come from the same family is hardly achievable, much less connecting absolutely any natural language.

1.1.2.3. Transfer approach

By the mid-1970s, researchers began to doubt the future of the interlingua approach and started to look at a more language-pair-specific and feasible approach. This led to the

[9] In comparison, in a direct translation system, to add a new language it was necessary to perform a morphological analysis of the additional language, write a bilingual (additional language > target language) dictionary lookup program, and write reordering rules (Hutchins & Somers, 1992, p. 72), then repeat these steps, but for translation of the target language to the additional language, as the direct approach is uni-directional.

third and most popular RBMT approach: the transfer approach. The transfer approach breaks the translation process into three parts: analysis, transfer, and generation. Essentially, a set of rules is established to break down the SL sentence, analyze it, and convert it into "abstract SL-oriented representations". Then, another set of transfer rules is used to convert the SL-oriented representations into "equivalent TL-oriented representations". After that, dictionaries are used to look up equivalents to the lexical units appearing in those representations and the TL sentence is generated from it (Hutchins, 1995, p. 3). While the basis of this approach closely resembles the interlingua approach, it should be noted that the transfer method requires intermediate representations that are entirely language-dependent. This means that every time a new language needs to be added to the system, more modules need to be created. Adding a third language to a bilingual system, for example, would require the creation of four new modules (compared to two in the interlingua approach). This number, however, increases with every additional language—adding a fourth language to a trilingual system would require the creation of six new modules (Hutchins & Somers, 1992, pp. 75-76).

While it may seem as if developing a transfer-based system requires more effort than developing an interlingua-based one, it is, in reality, quite the opposite. In fact, developing multiple language-dependent modules was easier than trying to develop a universal one. In an interlingua system, it is necessary that all ambiguities in the SL text be resolved before the translation can occur, whereas in a transfer system, only those inherent to the language in question needed to be removed (Hutchins, 1995, p. 3). Consequently, although it took more modules to add a language to a transfer system, it also took less time. The transfer approach thus became the preferred method and remained popular until the end of the 1980s. Nevertheless, MT did not reach large-scale commercial success during this period, giving way instead to CAT tools, the focus recommended in the ALPAC report.

1.1.3. Computer-aided translation tools

As researchers began to grasp the complexity of developing a truly autonomous and accurate RBMT system, the focus of research in the field of translation technologies shifted to developing CAT tools to assist translators instead. In a larger sense, CAT tools

could encompass any technologies capable of facilitating a translator's job, meaning that online dictionaries, grammar checkers, etc., could be included too (Bowker, 2002, p. 6). However, as these resources have become common to everyday computer users, we will only touch upon those that are specific to the task of translation and will highlight the differences between CAT and MT.

In the 1970s, researchers began looking into ways to increase translation productivity. Kay, for example, suggested an incremental approach to integrating computers in the translator's work, starting with tasks "not essentially related to translation" (2003, p. 226). The idea of translation memories (TM) and software (TM systems) to exploit them emerged and was implemented in the 1980s (Ferguson, 2019). The basis of the TM is to allow translators improve consistency and to eliminate repetitive tasks (LeBlanc, 2013, p. 6). The TM acts as a database that contains the source texts and their translated equivalents, aligned and divided into segments (usually pairs of sentences). The TM system is then able to automatically compare a new source text against the database of already translated texts it has compiled. The system presents the translator with matches, i.e., sentences that are identical (exact matches) or similar (fuzzy matches) to the sentence to be translated (Koehn, 2009, p. 242), and their stored translations, which the translator can decide to insert, edit and insert, or reject as required.

By the 1990s, CAT tools had become well established in the translation industry. In addition to using TM systems, language professionals were also familiarizing themselves with terminology management systems (TMS), terminology tools such as term extractors, and corpus-analysis tools such as concordancers. As tools to support translators in their work, they are all systems that rely heavily on human decision and judgment, both before the translation process and during the revision stage. In the case of TMs, for instance, the user needs to feed the TM during the building (or enriching) stage and has to edit the TL text in the revision stage, to ensure that the matches that have been incorporated into the translation are coherent with the rest of the text.

The fact that these systems require human direction could, in reality, be one of the reasons why translators favoured this technology over MT systems. As opposed to MT systems that automatically generate translations, TM systems "allow professional translators to be in charge of the decision-making" (Quah, 2006, p. 94). Indeed, the use of

MT in the workplace has long been frowned upon and remains a delicate matter that needs to be addressed. CAT tools, on the other hand, have been adopted not only by translation departments of corporations and big agencies since the 1980s, but also by the freelance community as of the late 1990s (García, 2006, p. 98).[10] This divergence in the perception of the technologies could be explained by the initial belief that MT systems were being developed in an attempt to replace human translators entirely (Marshman, 2014, p. 383). It has been observed that the reluctance to integrate MT into the workplace appears not to be rooted entirely in the quality of the MT systems and their product, but also (perhaps largely) in the human perception of such systems. This could partly be due to the "overselling" of the product, as we have seen previously with SMT and as pointed out by Castilho *et al*. In their 2017 paper, they stated that "[o]verselling a technology that is still in need of more research may cause negativity about MT, as already seen before with SMT systems […], when it was claimed that MT was producing 'near human quality' translations and that MT would 'steal translators' jobs', making translators 'merely post-editors of MT'. The hype that came with this euphoric presentation of SMT systems created a wave of discontent and suspicion among translators, that resulted in an 'us versus them' type of confrontation." (Castilho *et al*., 2017, p. 118).

1.1.4. Statistical machine translation

CAT tools were fully commercialized by the 1990s and had then already entered the workplace of not only in-house translators, but also freelancers. However, the dream of achieving MT was still being kept alive. In the late 1980s, after seeing how successful speech recognition was with the use of statistical methods, computer scientists started to research ways of using the same methods in MT (Koehn, 2010, p. 17).

Although the idea of using statistical methods for MT was first introduced in 1949 by Warren Weaver (Hutchins, 1995, p. 433), the idea of SMT was only born in the late 1980s, when a team from the Thomas J. Watson Research Center at IBM developed the first major SMT system. Known as project "Candide", the system was built using a corpus of texts from reports of Canadian parliamentary debates (Hutchins, 2001, p. 23). It

[10] This adoption was nevertheless not without its own challenges, as described, e.g., in García 2006, LeBlanc 2013, Marshman 2014. However, these are beyond the scope of this project.

should be noted that the biggest difference between RBMT and SMT is that the approach of SMT is not to generate one exact translation, but to generate thousands of possible translations and then to rank them based on their degree of correctness. Basically, the SMT approach begins with a statistical analysis of masses of human-translated, aligned texts. In that analysis, the system looks for patterns and regularities of lexical items' cooccurrence, both in a single language (to produce language models to predict likely word sequences) and across languages (to produce translation models to predict likely equivalents) (Koehn, 2010, p. 95). It then combines these models to predict the most likely translation for a given sentence (Forcada, 2017, p. 301).

Therefore, building a SMT system requires a large quantity of bilingual texts (a minimum of two million words for a specific domain or more for general language [Systran, n.d.]). Luckily, under the *Official Languages Act*, the proceedings of the Canadian parliament (called "the Hansard") have to be kept in both English and French, and are also available in computer-readable form. The IBM team was thus able to obtain "about 100 million words of English text and the corresponding French text from the Canadian government" (Brown *et al*., 1990, p. 82) to use as training data. The results were, as Hutchins put it (2001, p. 23), "surprisingly acceptable": almost half of the phrases translated in testing matched the phrases in the corpus, either word for word or with slight variations, or their wording was different but the meaning was correctly rendered nonetheless (Hutchins, 2001, p. 23).

Koehn recounts that research on SMT continued throughout the 1990s and peaked around the year 2000, with the increase, not only in computing power and data storage, but also in availability of digital texts, as a consequence of the growth of the Internet (2010, pp. 17-18). This led researchers to explore various models, essentially looking for different ways to divide the sentence into chunks to be processed by the SMT system. In the next sections, we will discuss the two best-known models for SMT systems: word-based and phrase-based. Although they are considered to be two different models, word-based and phrase-based SMT systems fundamentally work in the same way. Nevertheless, the size of the chunks influences how a system processes information and how it ultimately performs.

1.1.4.1. Word-based models

Both word-based models and phrase-based models rely on the same translation process, which is usually broken into three steps: 1) break the SL sentence into chunks; 2) find all the possible translations for each chunk; and 3) generate all possible TL sentences and find the most probable one (Geitgey, 2016). The main difference lies in the first step.

As Koehn (2010, p. 7) mentioned, IBM's original system was a word-based model that, as its name suggests, was translating words in isolation. This model can either use a dictionary to map words from one language to their equivalence in another, or, as in this case, can use translated sentence pairs for word-by-word alignment. The issue with this approach is that, while the model requires word-by-word alignment, what is mostly found in bilingual corpora is sentence-by-sentence alignment (Koehn, 2010, p. 88). Consequently, this makes step 2 hard to achieve.

Nevertheless, as explained by Koehn (2010, p. 95) the most likely translation (according to the data available) can still be found using Bayes' theorem of conditional probability. Bayes' theorem (sometimes referred to as "Bayes rule") is a mathematical formula for determining the probability of an event based on the knowledge of prior and posterior conditions. It essentially allows one to "fix or establish the validity of 'existing' or 'previous' beliefs in the face of best available 'new' evidence." (V V, 2016). This is relevant to MT because this principle can be applied to a statistical translation system in the form of an equation that optimizes the pairing of an SL word to a TL word.

Bayes' theorem of conditional probability is:

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)}$$

where H is a hypothesis and X evidence. We want to determine $P(H|X)$, the probability that the hypothesis H holds given the evidence X (Han *et al*., 2012, p. 350). If we want to apply this to SMT, we simply have to view the hypothesis as being the TL output and the evidence as being the SL input. In their original 1990 paper, the IBM team used "S"and "T" in their equation, as to describe their "source" and "target" sentences respectively. However, it should be noted that the concept of what constitutes the SL and the TL is different in mathematics and in translation (Koehn, 2010, p. 95). For this reason and for clarity, we have decided to explain Bayes' theorem in SMT with the equation that the

IBM team used in their 1993 paper (Brown *et al.*, 1993, p. 264) instead, where they cited French and English as examples for their SL and TL respectively. Thus, in a model where we are looking at translating from French to English, we have:

$$P(e|f) = \frac{P(f|e)\,P(e)}{P(f)}$$

where we are looking for $P(e|f)$, i. e. the probability of an English output sentence *e* given a French input *f*. If we look at the right side of this equation, we can consider $P(f|e)$ as "the probability that a translator, when presented with *e*, will produce *f* as his translation", while $P(e)$ is suggesting the order in which the words in the English sentence should be placed.

It goes without saying that in order to have the most accurate English translation (according to the hypothesis underlying SMT that the most likely translation is accurate), the probability $P(e|f)$ has to be maximized. Furthermore, since the denominator $P(f)$ is independent of *e*, we simply need to maximize the product $P(f|e)\,P(e)$ in order to obtain the maximized probability $P(e|f)$. Concretely, this gives us:

$$\underset{e}{\operatorname{argmax}}\, P(e|f) = \underset{e}{\operatorname{argmax}}\, P(f|e)\,P(e)$$

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, P(f|e)\,P(e)$$

where $\hat{e}$ is the estimated English output sentence for which the probability $P(e|f)$ is the greatest, i.e., the most likely translation (Brown *et al*., 1993, pp. 264-265).

Although Bayes' theorem helps to minimize the chance of error, it can only do so at the very last step of the SMT process, i.e., when the system is choosing the most likely translation among the possibilities generated. This means that if the data is incomplete or if errors were introduced in the system while breaking down the original sentence, there is no way to automatically correct such inaccuracies, and human postediting would be required. This, combined with the issues of not having word-by-word alignment readily available and of not having chunks that are always comparable in size from a language to another, prompted researchers to look at a new model where words could be grouped into bigger segments.

1.1.4.2. Phrase-based models

After the first explorations of word-based models, grouping words into phrases was considered to be the most logical way to obtain longer segments. This allowed phrase-based models to translate short word sequences as units, and, according to Koehn, this method remains the most effective approach for SMT (2010, p. 127). Indeed, the phrase-based model not only addresses the problem of having possible discrepancies in word-phrase alignment, it also resolves some translation ambiguities.

A standard model for phrase-based SMT consists of three phases: segmentation, translation, and reordering. In the segmentation phase, the sentence is broken down into multiword units that are not necessarily determined by linguistic rules or syntatic theories (Koehn, 2010, p. 128), but are rather seeking to provide a useful context. This will allow the system, in the subsequent translation phase, to better determine which translation is the most appropriate for a word, given its co-occurrence (i.e., the other words surrounding it). It greatly increases the accuracy of translation.

In the translation phase, each phrase is translated and each translation is ranked according to Bayes' theorem, the most probable translation receiving the highest score. After all the phrases have been translated, they can then be reordered using a distance-based reordering model (Koehn, 2010, p. 129) so as to correspond to the TL syntax.

Moreover, phrase-based SMT also offers a third benefit not found in word-based SMT: if the training data is vast enough, the system may be able to learn or "memorize" longer and longer phrases (or even sentences), which further accelerates the translation process (Koehn, 2010, p. 128)

Overall, SMT performs much better than RBMT if given sufficient training data (Isabelle *et al.*, 2007). Its strength also relies on the fact that compared to RBMT, SMT requires much less development effort, as there is no need to come up with complex and exhaustive rules. However, SMT systems still have their limitations, as they are still complicated to build and maintain. For rarer language combinations, they also oftentimes require the use of a pivot language (i.e., an intermediary language, often English, through which the translation is mediated), as extensive bilingual data is not always available for every language pair of interest. Most importantly, "humans are still needed to build and tweak the multi-step statistical model" (Geitgey, 2016).

1.1.5. Neural machine translation

While SMT clearly has many advantages over its RBMT predecessor, some concerns about quality (e.g., inconsistency in terminology or vocabulary use) and logistics (e.g., the reliance on the availability of suitable corpora) remained. In 2014, Sutskever and his team introduced a new approach to MT using RNNs, which revolutionalized how MT was viewed as a problem.

As Geitgey explained it, RNNs are a tweaked, more sophisticated version of neural networks that have a *stateful* model. Regular neural networks are a "generic machine learning algorithm that takes in a list of numbers and calculates a result" (Geitgey, 2016). The model is said to be "stateless", meaning that it does not have a memory. Consequently, the neural network will always return the same output when the same input is entered, as it is unable to detect patterns in data over time. In a translation context, an MT system with this stateless model would still require human assistance for maintenance and—most of all—updating, which would not be a significant upgrade compared to an SMT system. With the use of RNNs, however, the system retains a memory of previous calculations and can make predictions while considering what it has most recently seen. In other words, the system can update itself every time it is used, and previous calculations can influence future outputs.

Sutskever's work was exceptional because a classic (also called "vanilla") RNN is usually able to compute a sequence of outputs given a sequence of inputs, but this mapping is usually done when the alignment between the input and the output are known ahead of time (Sutskever *et al.*, 2014, p. 3). This is problematic for MT, as the input and output are more often than not variable in length and have "non-monotonic relationships," meaning that if a sentence is aligned with its translation, the relationships between the words in the sentence will not always be parallel (see Figure 1).

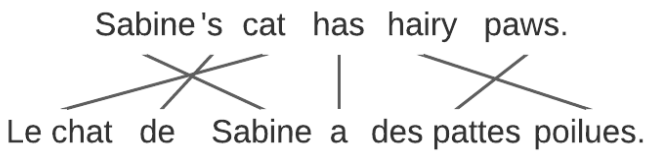

Figure 1 – Example of a non-monotonic relationship where alignments are crossing between parallel sentences

While these classic RNNs are very good at solving sequence-to-sequence problems, their weakness lies in the fact that they have only a limited window in which they can connect information to build context. This leads to what are referred to as long-term dependency problems, in which the choice of a TL item is conditioned by an SL item that is at some distance from the item currently being translated. In an MT system where translated words are generated sequentially, classic RNNs might be able to guess the next word based on previous words in a short sentence, but they will most likely struggle to retrieve information the further back or forward they have to look, and the longer the sentence becomes. To address this issue, Sutskever et al. proposed a refinement to the classic RNN architecture, called Long Short-Term Memory (LSTM). LSTMs were first introduced in 1997 (Hochreiter & Schmidhuber, 1997) and are designed to avoid long-term dependency problems. By having four layers of neural networks (i.e., deep neural networks) instead of one, as in the classic RNN, Sutskever's LSTM RNNs were able to do well on long sentences, with no degradation on sentences with less than 35 words (Sutskever *et al.*, 2014, p. 7). His model uses two different LSTM RNNs, where the first RNN encodes a source sentence and the second one decodes a target sentence. Sutkever's work with LSTM RNNs can be considered pioneering in NMT, as his team was among the first to leverage deep learning in MT.

The same year, another team developed a model inspired by Sutskever's sequence-to-sequence approach and named it the "RNN Encoder-Decoder" model. This also consisted of two RNNs, but with a simplified architecture (Cho *et al*., 2014b, p. 3). Cho's team's RNN Encoder-Decoder was easier to develop and implement than Sutskever's model, and was able to learn from parallel corpora to translate sentences from English to French. Moreover, they presented a new hidden unit in their model that is capable of adaptively remembering or forgetting information, further reducing the human maintenance that was previously needed for updating corpora as language evolves, for example.

Following Cho's work, RNNs became the standard for MT because they "are useful any time you want to learn patterns in data" and "human language is just one big, complicated pattern" (Geitgey, 2016). Since the development of RNN-based systems,

two other approaches to NMT have also been explored: CNN-based and attention-based. These architectures will be discussed in more detail in section 1.1.5.1.2 and 1.1.5.1.3.

Overall, NMT systems seem to render better translations than SMT systems and many researchers have used case studies and BLEU scores to demonstrate this (Bentivogli *et al*., 2016). Nevertheless, there are still risks associated with the technology being used in a professional setting, including but not limited to translations that are fluent-sounding but semantically inappropriate and mistranslation of out-of-domain texts or vocabulary (i.e., texts that are from a different domain than the texts used for training the system, or vocabulary absent from that training data) (Koehn & Knowles, 2017, p. 33). Indeed, since a RNN acts as a black box (i.e., a device that produces an output based on an input without revealing any information about its internal working). Forcada (2017, p. 301) has stated that errors in NMT are much harder to trace because they cannot easily be traced back to the bilingual corpus used to train the system. With RBMT and SMT systems, when an error is detected, its source can generally be identified and it can typically be corrected by adjusting or adding rules (for RBMT), or by adding correct examples to the training data (for SMT).

According to Google, NMT has three main weaknesses: "its slower training and inference speed, ineffectiveness in dealing with rare words, and sometimes failure to translate all words in the source sentence" (Wu *et al*., 2016, p. 2). Consequently, considerable work on NMT has been focused on improving these three areas.

1.1.5.1. Neural machine translation architectures

In this section, we will discuss specifically the three most prominent architectures for NMT systems: RNN-based systems, CNN-based systems, and attention-based systems. These three architectures are considered to be state-of-the-art as of the writing of this thesis, and are currently being used in the most popular systems and deployed on the most popular platforms (Google Translate, DeepL Translator, Facebook, Amazon Translate, Azure Translator, etc.). In the following sections, we will elaborate on their distinctive characteristics and discuss their strengths and weaknesses at a more technical level.

1.1.5.1.1. Traditional recurrent neural network-based systems

RNN-based NMT is now essentially achieved through the Encoder-Decoder model presented by Cho *et al.* (2014b), where an encoder RNN "consumes" the source sentence and a decoder RNN produces the target sentence (Figure 2).

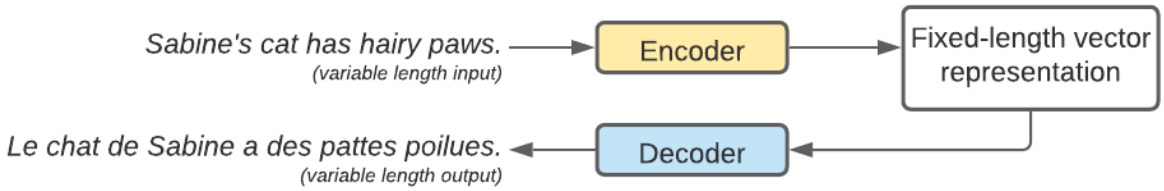

Figure 2 – Encoder-decoder structure

As Geitgey illustrated in his Machine Learning Series (2016), the encoding and decoding process can each be summarized in three steps: feeding of an input into the RNN, processing of the information through the hidden state (or black box), and generation of an output. During the encoding, the input (which consists of a variable-length sentence that has been segmented into a sequence of units of information [tokens] during a pre-processing stage called tokenization),[11] is fed into the first RNN. The hidden state is the element that allows NMT models to learn on their own and improve over time. It is essentially a loop that retains information from previous inquiries and uses it in subsequent calculations. Neural networks, in general, can only process numbers and arithmetic operations. Furthermore, they need data that has a fixed length in order to perform batch operations. This is why the text (i.e., the input) needs to be converted into numbers at this point: this step is called word embedding (or vectorization).[12] The output of this step is the fixed-length encoded sentence: a series of unique measurements (or numbers) that represents one of the tokens that has been input into the neural network. Naturally, this series of numbers is not our desired result—we want to obtain a new sentence in the target language. This is where the second RNN comes into play.

During the decoding, the input becomes the series of unique numbers produced during encoding. This series of numbers is then put through the hidden state of the second

[11] We will further detail the process of tokenization in section 2.2.2.2.
[12] Word embedding will be further explained in section 2.2.2.3.

RNN, which reversely converts the numbers into units of information, but this time, in the target language. This notion of conversion from units of information to numbers or from numbers to units of information is illustrated in Figure 3.

Essentially, the first RNN is an encoder that reads each unit of information from the input sequence, in a sequential manner, and converts them into a list of unique measurements. As it calculates the input data it receives, the hidden state of the RNN changes and saves a summary of the entire input sequence. The second RNN then acts as a decoder and generates the output sequence by predicting the next unit, given the new hidden state (Cho *et al*., 2014b, p. 2).

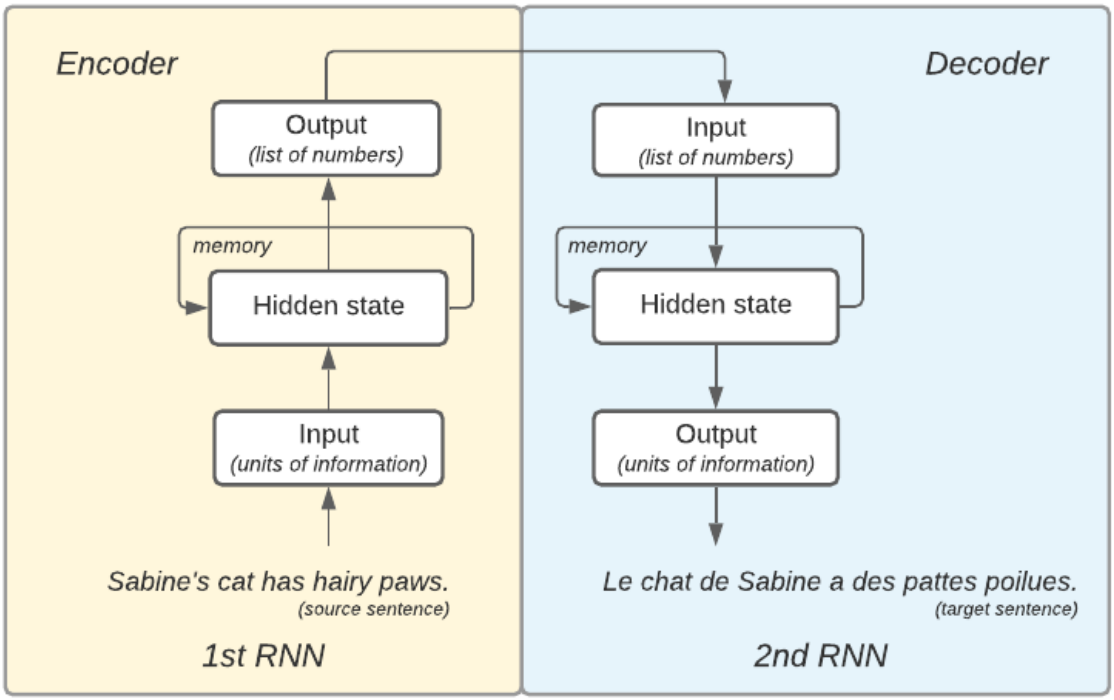

Figure 3 – Simplified RNN-based model architecture

Cho's RNN-based model differs from Sutskever (see 1.1.5) because his team used gated recurent units (GRUs) instead of LSTMs. GRUs are similar to LSTMs, but are easier to develop and implement. While these two units perform comparatively in MT tasks (Chung *et al.*, 2014), GRUs are often used because of their less complex structure.

1.1.5.1.2. Convolutional neural network-based systems

While some researchers have been looking at ways to enhance RNN-based NMT systems to overcome some of the challenges mentioned at the beginning of this section (see also 0), others have decided to adopt a whole new approach to NMT using convolutional neural networks (CNNs). RNNs became the standard for MT because they are able to create a fixed length vector from a variable length input, whereas CNNs usually only create representations for fixed-size contexts, making them useful for tasks such as image recognition (Bergen & Wagner, 2015), face recognition (Lawrence *et al.*, 1997), sentence classification (Kim, 2014) and document recognition (LeCun *et al.*, 1998. However in 2017, Gehring and the Facebook AI Research (FAIR) team published a paper on a CNN-based approach to MT. Although Gehring and his team were not the first ones to use CNNs in MT, they were the first to show that CNNs applied to machine translation could outperform RNNs (Yarats *et al*., 2017).

Gehring's model (presented in Figure 4) follows the encoder-decoder approach, but uses two CNN encoders and one LSTM RNN decoder (the same as the one found in Sutskever's model [1.1.5]). Gehring and his team argue that RNNs are not the best fit for NMT as "modern" machine learning leverages the use of graphics processing units (GPUs). GPUs have a highly parallel structure and are good at running the same task on different data, in parallel (data parallelism). Using RNNs for MT is restrictive, as each word must be processed in a sequential manner, meaning that "each word must wait until the network is done with the previous word" (Yarats *et al*., 2017). Conversely, CNNs are able to compute all the features of the source sentence simultaneously, thus avoiding having "the first word […] over-processed and the last word […] transformed only once." (Gehring *et al.*, 2017).

The first CNN in Gehring's model is an encoder that calculates attention scores (also referred to as *weights*),[13] while the second CNN, also used as an encoder, calculates a conditional input. Gehring and his team used a *dot-product style attention* mechanism in their model. The attention mechanism gives the decoder direct access to the encoder so it can focus on a particular section of the source sentence at a time. This is particularly

[13] We will dive deeper into attention mechanisms in the next section on attention-based NMT

useful as the sentences to be encoded become longer and longer, as it allows the system to handle partial sentences rather than being dependent on the calculations from the entire sentence, which is long, complex and demanding as the sentences get longer. By adding attention to the system, as the system processes one word at a time, it can focus on different parts of the sentence to build the context it needs to achieve an accurate analysis of the word it is processing at that moment. It eliminates the need to consider the sentence as a whole, which can be computationally demanding in longer sentences.

An attention score is produced for each token of the source sentence and those scores (also referred to as *dot products*) are computed into a series of probability distributions, determining the most likely sequence of translated outputs.

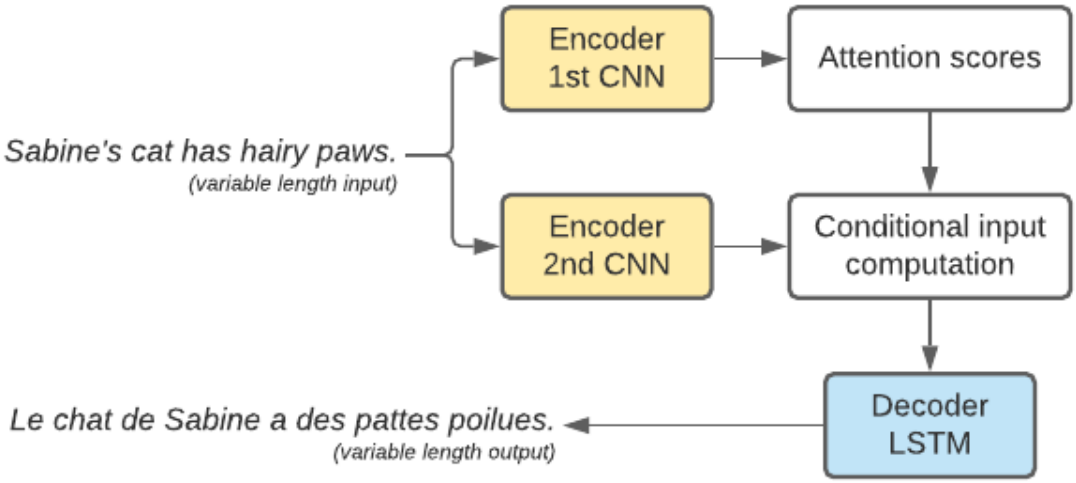

Figure 4 – Simplified Gehring model architecture (embedding layers omitted)

The conditional input is a weighted sum of the attention scores and the source embeddings.[14] What is obtained from both CNNs is then put through a LSTM decoder. The decoder computes a new hidden state, which contains information about the previous hidden state, and ultimately produces a target translation.

Gehring and his team found that their CNN encoders' performance was comparable, or superior, to LSTM encoders when applied to tasks such as WMT'16 English-Romanian, WMT'14 English-French, and WMT'15 English-German. They also reduced translation time with their CNN encoder, stating that models which utilize their

[14] Word embedding will be further discussed in section 2.2.2.3.

encoder are able to translate twice as fast as strong baselines with RNN encoders (Gehring *et al*., 2017, p. 8).

1.1.5.1.3. Attention-based systems

Though CNN-based MT significantly reduced processing time and solved some of the context problems that RNN-based MT had, some issues still remained, driving researchers to explore a new, less complex, architecture.

Attention mechanisms are essential components that are at the core of the best performing encoder-decoder models, which is why in 2017, a team at Google introduced a new NMT model that fully relies on attention and does not require recurrence or convolutions. Vaswani and his team proposed a model, named *Transformer*, that produced higher quality translation, while requiring less time to train (Vaswani *et al*., 2017). In many of the models introduced prior to the *Transformer*, attention mechanisms were not at the core of the architecture, but were instead used to complement either an RNN or a CNN.

The main issue with RNN-based NMT, as raised by researchers, is that models are complicated to build and require more and more computational power to run the longer a sentence becomes, due to memory constraints (Vaswani *et al*., 2017, p. 2). The sequential nature of RNNs does not allow for parallelization (that is, the translation process is word by word, the system cannot translate multiple words at a time), which tends to lengthen the translation process. Attention mechanisms have made it possible to reduce sequential computation by allowing the system to focus on specific sections of the sentence at a time, eliminating the need to process the entire sentence all at once. They have also allowed for modeling of dependencies between words regardless of their distance from one another.

The Transformer has a similar architecture to the other models presented above, except that the encoder and the decoder are each made up of six identical layers (as represented in Figure 5).

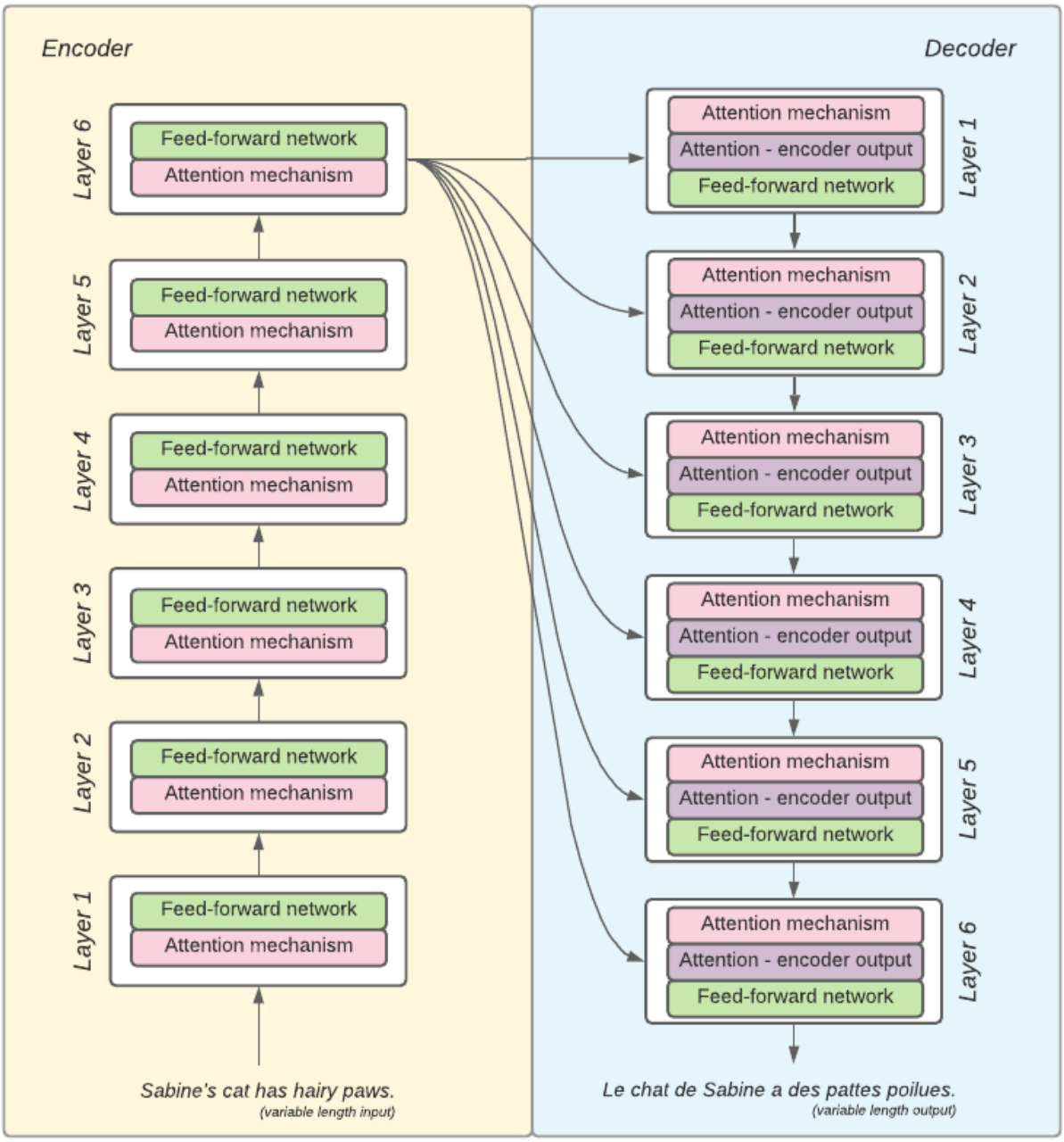

Figure 5 – Simplified Transformer model architecture

In the encoder, the layers are made up of an attention mechanism called "self-attention" and of a feed-forward network. Self-attention is what allows the model to build context: as it processes each word from the input sequence, self-attention allows the model to consider (or pay attention to) other words in the sequence in order to get a better understanding of the word it is processing at that moment. Determining accurate semantics will, in turn, allow for a more accurate encoding of that word. This can be

compared to the hidden state we saw in RNN-based models. The feed-forward network is simply a neural network that does not present a loop (as opposed to RNNs). As its name states, information in a feed-forward network only moves forward.

In the decoder, in addition to the self-attention mechanism and the feed-forward network, Vaswani and his team added a third sub-layer ("Attention – encoder output" in Figure 5) that performs attention. This third sub-layer allows the decoder to focus on specific parts of the output of the encoder, hence the word attention.

A particular characteristic of the Transformer is that its structure allows developers to visualize which parts of the sentence the encoder is focusing on at each step of the encoding. This made it possible for them to gain insight into how the system deals with a well-known task in natural language processing (NLP): coreference resolution (Uszkoreit, 2017). In translation terms, this would be the ability of the system to accurately translate deixes, and from preliminary tests done by Vaswani and his team, the Transformer was able to accurately translate the pronoun "it" in sentences that Google Translate did not successfully translate. Therefore, in theory, the Transformer should be stronger than an RNN-based system at translating such items.

1.1.5.1.4. Hybrid NMT systems

Some studies have shown that NMT systems are fluent (grammatically correct and idiomatic), but sometimes inaccurate (Tu *et al.*, 2017, p. 2), while SMT systems are generally able to produce translations that convey the correct meaning of the source. As SMT and NMT systems have their own strengths and weaknesses, researchers have developed hybrid models, leveraging the strong points of the two different architectures.

Some, for example, added SMT features to an NMT system. This was the case for a team from Soochow University and Huawei Technologies proposed a hybrid model consisting of an SMT model in an NMT framework (Wang *et al.*, 2016). In their model, at the decoding stage, an SMT model would suggest additional translation candidates based on the decoding made by the NMT system. Overall, they found that without the SMT recommendations, their system achieved a lower BLEU score.

Another team from Baidu targeted specific problems known to NMT and developed a model combining an RNN-based model and a log-linear SMT model (the

log-linear model being the dominant framework for SMT). They incorporate three SMT features to their NMT framework: the translation model, the word reward feature, and the language model. The translation model relied on conventional PBSMT approach, the word reward feature "controls the length of the translation" (He *et al*., 2016, p. 151) and causes the decoder to prefer long translations (p. 152), while the language model aimed to enhance fluency. With this new model, they achieved higher BLEU scores on Chinese to English translation tasks and were able to improve the translation quality overall (He *et al.*, 2016).

Others, such as Torregrosa *et al.* (2019) took a different route used the information (rules and dictionaries) contained in an RBMT system to potentially tweak an NMT system, therefore resulting in a hybrid RBMT-NMT system. Torregrosa and his team added morphological information to the source language and found that this method was as effective as using subword[15] units in a low-resource setting, that is, when the availability of parallel corpora may be limited for a certain language pair.

These are only some of the examples seen in terms of hybrid systems involving a neural component. All in all, there have been efforts to increase the quality of translations produced by NMT systems by focusing on known weaknesses of the neural architecture. However, there is still work to do to understand the specific weaknesses of each architecture, which is why our project aims to contribute to this understanding.

1.2. Evaluation of machine translation

As the number of MT systems continues to increase, evaluation methods had to be developed to enable researchers to track progress and to compare the systems to one another. Researchers had to look for ways for human judges to assess MT in an objective manner, while still considering the limits of what can be expected of MT systems. Hutchins and Somers (1992, p. 161) highlighted that, in the early years of MT, most evaluation was carried out either by non-professionals who lacked knowledge about MT, or by developers who—perhaps because they wanted to show their system's strengths—would carefully select their evaluation data (i.e., use sentences or phrases they knew their

[15] Subwords allow for an "open vocabulary", allowing the system to process words it has not previously seen in its training data. These notions will be further discussed in section 2.2.2 of our Methodology.

system was able to tackle), ultimately leading to misleading results. Although human evaluation of MT is very valuable and provides insight into many aspects of translation, including the three elements cited by Hutchins and Somers—accuracy, clarity, and style (1992, p. 163)—it remains a method that is expensive, time-consuming, and, to this day, not always objective. To cite only one example, language professionals who perceive MT as a threat can sometimes be biased when told they are evaluating a machine-translated text (Roturier, 2006, p. 81). Consequently, researchers started to explore new methods of automatic scoring that could reduce time and cost, while still covering accuracy, clarity, and style. Out of these new methods came BLEU, now considered as the standard scoring method for MT.

While BLEU and its like offer quick and comparable scores, the method has its weaknesses, some of which have been revealed as MT has evolved (Callison-Burch *et al.*, 2006; Novikova *et al.*, 2017; Reiter, 2018). To address these problems, researchers started to go back to a more traditional method of evaluation, involving humans once again. However, asking humans to manually score MT systems while maintaining a level of consistency and objectivity meant that researchers needed to better define what they wanted to evaluate in a system.

In the following sections, we will present the strengths and weaknesses of the BLEU score, briefly present some additional approaches to evaluation using human judgment, and, finally, explain what a challenge set is and why challenge sets are being revisited in NMT evaluation.

1.2.1. Automatic scoring

Throughout the years, many methods of automatic scorings (sometimes referred to as *automatic metrics*) were developed for MT. BLEU is only one of them,[16] but perhaps the best known in the industry. Other methods are often based on BLEU (e.g., the National Institute of Standards and Technology [NIST]) or try to tackle issues not covered by BLEU (e.g., Metric for Evaluation of Translation with Explicit ORdering [METEOR]).

[16] For a more complete review of MT evaluation methods, consult Kit & Wong (2014) or Koehn & Monz (2006), among others.

We will focus on BLEU, as it is both the current industry standard and the scoring method we used during training of our systems (see section 2.2.3 for more details).

In 2002, Papineni, Roukos, Ward & Zhu, these researchers from the IBM T. J. Watson Research Center presented a method for automated evaluation of MT that they named the **bil**ingual **e**valuation **u**nderstudy, better known as BLEU. The idea behind BLEU was to release MT researchers from the evaluation bottleneck that was slowing down MT progress and provide them with "an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation" (Papineni *et al.*, 2002, p. 1).

Papineni and his team believed that "[t]he closer a machine translation is to a professional human translation, the better it is," which is why their MT evaluation system relies on two elements: "1. a numerical "translation closeness" metric; 2. a corpus of good quality human reference translations" (Papineni *et al.*, 2002, p. 1). BLEU compares the candidate translation to available reference translations and assigns a score to the candidate translation that ranges from 0 to 1. A score of 1 means that the candidate translation is identical to a reference translation (meaning that even human translations will not necessarily receive a perfect score); the more reference translations there are per sentence, the higher the score, since the more matches there can be between candidates and references. After comparing BLEU scores to human evaluations, Papineni *et al.* found that they correlated highly, stating that "BLEU tracks human judgment well" (Papineni et al., 2002, p. 7).

While BLEU scores are very useful for frequent and repetitive tasks such as monitoring a system's performance during its development stage, or for system comparisons, the method still as it does not provide enough information to assist researchers in their quest for "genuine improvement in translation quality" (Callison-Burch *et al.*, 2006, p. 1). In their research, for example, the authors found that improvements in BLEU scores do not correlate with improvements in translation quality (2006, p. 1), thus questioning the true accuracy of the technique. They thus questioned the method's robustness and clarified that the BLEU scoring system should only be used in specific settings, for specific ends: BLEU alone is not sufficient for "comparing systems which employ radically different strategies," i.e., different architectures.

In the next sections, we will give an overview of some of the human-driven approaches to MT evaluation and then will present the challenge set method, as it is the approach we will be adopting.

1.2.2. Human evaluation of machine translation

As progress in the field of MT largely relies on representative evaluation of the systems developed, researchers started to realize that using an automated scoring method might not be enough to evaluate MT and all of its aspects. They started to go back to the more traditional approach of involving human judges, but established more specific criteria for evaluation. In this section, we will give a high-level overview of the methods used and—although this will not be an exhaustive portrait of human-driven evaluation of MT—will show what each of them can bring out.

There are multiple approaches to evaluation of MT assisted by humans. As part of our research, we will mostly discuss the methods that focus on the quality of the output (as opposed to methods evaluating the source for translatability, for example, or on the robustness or usability of the systems), as we are interested in MT from an end-user perspective. Two common human evaluation metrics used for MT are fluency and adequacy (White *et al.*, 1994, p. 196; Koehn & Monz, 2006, p. 102). Fluency measures whether a translation is fluent (i.e., if the translation would sound natural to a native speaker), while adequacy measures whether the translation conveys the correct meaning of the source (Snover *et al.*, 2009, p. 259). These metrics are often measured together on a 5 or 7-point scale, and their average is used as an overall translation quality score (p. 259). However, this method still relies on human judgment, and thus still entails some subjectivity.

Other more process-based and objective strategies have been suggested; for example, the use of eye tracking on MT outputs (Doherty *et al.*, 2010) to measure the average gaze time (i.e., when a reader gazes over an area of interest) and fixation count (i.e., when a reader focuses on a specific part of the sentence), as indicators of cognitive effort (i.e., the effort expended by the reader to understand the translation). These measures were found to be higher in "bad" translations (suggesting that poor translations require more cognitive effort) than they were in translations that were rated as excellent

in an earlier human evaluation step (Doherty *et al.*, 2010, p. 1). Although this method involving human evaluators is objective, is it also very costly. Moreover, cognitive effort is a fairly abstract measure of quality that may be difficult for some evaluators (and decision-makers) to relate to their own situation. Methods targeting more immediately obvious effects of translation quality (e.g., on user comprehension), and that are also less costly, have been explored. One of these is gap filling (GF) (Forcada *et al.*, 2018, p. 192), in which certain words are removed from the reference translation and human evaluators are asked to fill those gaps, using the translation produced by the MT system as an aid. The approach measures the proportion of gaps that are successfully filled (Forcada *et al.*, 2018, p. 194). While such methods can be useful to evaluate MT-produced texts for gisting, neither this approach nor eye-tracking are likely to be precise enough to allow targeting of specific issues that will allow developers to improve the quality of MT, particularly if the intent is to produce publishable-quality texts.

For this reason, researchers have started to look into evaluation methods that are more tailored to the specific problems in MT (and NMT in particular) that they would like to tackle. Popović, for example, presented different approaches to classifying and analyzing MT-related errors found in outputs, arguing that manual evaluation of MT using human evaluators suffers when there is no automatic error analysis (Popović, 2018, p. 129; 2021, p. 163). Error analysis typically begins with error classification. In an error classification task, annotators are asked to mark each erroneous word in a translation and to assign a corresponding tag to it (p. 131). The error categories (or typologies) are defined beforehand to maintain consistency, and as a reference, annotators also have access to either the original source text, a reference translation, or both.

Another approach for error analysis is to use contrastive sets. Tang *et al.*, evaluated the three most prominent NMT architectures on subject-verb agreement and long-range dependencies (Tang *et al.*, 2018). They used sets of contrastive translations to analyze specific errors and paired each contrastive variant containing an error to a human reference translation, free of errors. They then had their NMT model assign probability scores (i.e., scores that would generally determine the most likely translation) to the sentences in the pair, and if the model assigned the highest score to the correct translation, they considered it a correct decision. Contrastive evaluation allowed

developers to target specific weaknesses of a system rather then rate their overall quality. They are an interesting complementary evaluation method to BLEU scores.

This motivated researchers to test out a related approach, where, instead of looking at the errors their system was producing, they started by listing potential errors and testing their system to see if the errors would occur. This "challenge set" approach's main advantage over the linguistic categories previously mentioned is that it allowed for a controlled distribution and frequency of the phenomena of interest (Popović, 2018, p. 151). In the next section, we will dive deeper into the characteristics of challenge sets.

1.2.3. Challenge sets

In the early 90s, challenge sets were popular as they were practical for "probing syntactic competence of grammar-based MT" (Popović & Castilho, 2019). With SMT, researchers started to opt for evaluation using standard natural test sets as the prevailing evaluation technique instead (i.e., with naturally occurring data). However, with the emergence of NMT, some researchers went back with the challenge set approach that allowed for insight into some particular phenomena.

Challenge sets, also known as test suites, used on NMT systems were first introduced by Isabelle, Cherry, & Foster in 2017 as a "new evaluation methodology (…) designed using expert linguistic knowledge to probe an MT system's capabilities" (Isabelle *et al.*, 2017, p. 1). The method is straightforward: Construct sentences that each contain a specific linguistic phenomenon. Then, have humans manually evaluate the resulting translations by answering a yes or no question concerning the correctness of that particular phenomenon (e.g., "Is subject-verb agrement correct?", "Does the flagged adjective agree correctly with its subject?", etc. [Isabelle *et al.*, 2017, p. 13]). This allowed for "a more fine-grained picture of the strengths of neural systems," in addition to providing insight on which aspects of the language remained unrealistic for a machine to tackle. Isabelle & Kuhn (2018) developed a corresponding French to English challenge set in subsequent work.

Related to Isabelle, Cherry, & Foster's research, Koehn and Knowles (2017) also explored challenges for NMT, only they studied six challenges to give empirical results, comparing NMT to SMT, rather than to establish a challenge set. They opted for non-

linguistic-based challenges and looked at elements such as domain mismatch and the amount of training data. Their findings provided more insight into how the NMT model could be improved, as they discovered problems linked to the training process rather than the performance.

Although challenge sets prove to be useful for NMT evaluation, they still have weaknesses and, according to Popović and Castilho (2019, p. 1), should not be used on their own. Challenge sets allow us to look at a very particular phenomenon in a sentence but do not provide input on the quality of the translation as a whole. In fact, a sentence that contains gross errors could still be evaluated as having successfully passed the test as long as the challenging element of the sentence was correctly translated. Popović and Castilho noted that challenge sets "do not reflect the statistical distribution of phenomena encountered in naturally occurring data" (Popović and Castilho, 2019, p. 1). This is why they suggest using both a challenge set and natural test sets to test a system.

While we will only be looking at the challenge set approach in our paper, we recognize these weaknesses and will take them into account when we look at our results. They will help guide our analysis and will allow us to focus on some of the things that are identified as relevant differences between NMT models in terms of performance. We are also aware of the possible subjectivity in judging correctness of MT systems in a challenge set, particularly relevant here as (due to time and resource constraints inherent in a Master's-level project) a single perspective is represented in creating reference translations and in evaluating system performance. Our reference translations are nevertheless presented as a point of reference, but as only one of many possible and acceptable solutions. Other anticipated—and unanticipated—solutions are still considered as acceptable if they meet the standard set for the project. We are using a standard based on our own training and experience, of what we believe a Canadian-trained professional translator striving for publishable quality would feel the need to post-edit. While this necessarily entails personal judgment, we believe that it does allow us to make decisions that are as realistic and consistent as possible in the circumstances.

Another suggestion would be to complement challenge sets with document-level human evaluations of MT. Castilho, Popović, and Way have stated that what constitutes "document-level" evaluation is still unclear, as such evaluation could refer "to pairs of

consecutive sentences, to a paragraph, or even to whole chapter", but they emphasized the need for context-aware MT evaluation, as more and more developers work on discourse-level MT systems and need the appropriate evaluation methods to improve their systems (Castilho *et al*., 2020, p. 3735).

Document-level evaluation of MT is a great complementary evaluation method to the traditional BLEU score and to challenge sets because it extends on adequacy and fluency and helps identify where the problems are located. It achieves a good balance between BLEU, a method that is somewhat imprecise but holistic, and challenge sets, which are very precise, but perhaps too focused. We suggest going through an incremental evaluation process where one could first use a BLEU score to benchmark a system against others that have already been published. If the system is comparable to previously published ones, one could then use a challenge set to evaluate the MT system at the sentence level. Finally, if the results are deemed good, one could then proceed with using a document-level evaluation method.

In Chapter 2, we will be outlining the work we put into training our systems and will briefly introduce our datasets and our evaluation metrics. In Chapter 3, we will dive into the creation of our challenge set, from the reference manuals and other resources on which we based it, to the detailed justifications of what constitute a "correct translation".

Chapter 2: Methodology

In this chapter, we will discuss the various steps to training and developing our MT systems. Before getting into this discussion, however, we will briefly describe three systems we used as points of comparison for our experimental NMT systems: one hybrid SMT system, and two large-scale online NMT systems. This comparison helps to highlight the overall performance of the experimental systems in relation to the "standard" alternatives available and to bring out some of their more unusual features. We will then present the entire process of selecting our training data, preparing that data, and training each of our NMT models.

2.1. Portage, Google Translate, and DeepL Translator

In this section, we give a brief overview of the three systems we used as points of comparison. For the hybrid SMT system, we used Portage, a proprietary system developed by the NRC that can be used commercially or to support other NRC projects. In fact, an early version of Portage is still available, at the time of writing, on the Government of Canada's internal network, and is intended to be used as a gisting/comprehension tool for public servants.

As of Portage II-3.0, the NRC has incorporated deep learning into their state-of-the-art SMT model, which reportedly substantially improved the system's performance. Portage II-3.0 features a neural network joint model (NNJM), which helps capture long-term cross-lingual dependencies (Devlin *et al*., 2014; Joty *et al.*, 2017, p. 165). The NNJM is added to the target language model and it "[e]stimates the probability of a target word given its previous word history and a source context window" (Banchs, 2016, slide 160). In other words, it takes into account the surrounding context of the word being translated (i.e., in the source), in addition to considering the words that have already been translated (i.e., in the target) and that are preceeding it. In their research, Devlin et al. reported that this new component allowed for an increase in BLEU score.

In addition to Portage, we also used to two large-scale NMT systems: Google Translate and DeepL Translator, both of which are available online. We chose these systems because they are well known and frequently updated, giving us a more accurate

representation of the state of NMT at the time of our analysis.[17] These systems were used during the development of our challenge set (as discussed in 3.2), as well as for comparison of our results against theirs.

Google Translate (hereafter Google), originally an SMT system, became an NMT model in 2016 (Le & Schuster, 2016). Like Sutskever's model (1.1.5), their model uses LSTMs but has eight layers in the encoder and eight layers in the decoder (Wu *et al.*, 2016, p. 1). They used BLEU to evaluate their model, but also had bilingual human evaluators rate the overall quality of their translations by presenting them with two translations, side-by-side, for a given source sentence (p. 14). They found that their neural model "approaches the accuracy achieved by average bilingual human translators on some of [their] test sets" (p. 20). Compared to the PBSMT model they had prior to this neural model, they observed "roughly a 60% reduction in translation errors on several popular language pairs" (p. 20).

DeepL Translator (hereafter DeepL) is also an online NMT system, launched in August 2017, with seven languages at the time, but 24 as of the writing of this thesis. DeepL was developed by the same German company that operates Linguee, a well-known online concordancer. Linguee uses webcrawlers to gather professionally translated texts that are publicly available in two or more languages. These texts range from translated texts on a company's website to government-official documents from the European Union or from Canada, which makes Linguee's database rich in highly specialized tests (Lardinois, 2010). The texts are aligned on a sentence level and the quality of the translations is evaluated using an in-house automated algorithm to determine whether the translation is "good enough" (Lardinois, 2010). According to Lardinois, the co-founder of Linguee, Leonard Finke, has said that this is what makes the tool excel at handling polysemous words and idiomatic expressions. DeepL is said to rely on the bilingual, human-generated, dictionaries that were built for Linguee (Ziganshina *et al.*, 2021, p. 4), which may explain why in some of the studies comparing DeepL to Google, DeepL tends to achieve higher scores (Isabelle & Kuhn, 2018, p. 7). Comparisons of our results against these two systems will not only provide us with a

[17] Our experimental systems may be outdated given that we used models available when we first started drafting this thesis and developing our challenge set.

benchmark, but will also allow us to use these two systems' handling of specific difficulties as a basis of comparison for our systems.

2.2. NMT system development

Before we get into the specifics of training each of our selected models, we will describe the preliminary steps that are required to develop any NMT system. In preparing our experimental systems, in order to observe and focus on the effects of the system architectures, we ensured that the systems were as comparable as possible in other ways, including the hardware (insofar as possible), training data and pre-processing. In the following section, we will present these common steps.

2.2.1. Training data

As mentioned previously, our challenge-focused evaluation method mainly relies on translation difficulties discussed in translation pedagogy, writing for translation, translatability, and computers and language. These issues are thus not limited to a given domain or sub-language, but rather identified as widely relevant for translators, writers, and other language professionals. For this reason, it was important to train our systems using datasets that are not domain-specific. We also required large corpora for the English-French language pair that were freely available and accessible for research purposes, to ensure that the systems had sufficient data to function relatively well.

Based on these critera and taking into account the recommendations made by our NRC collaborators, we selected the four datasets described below.

- commoncrawl.fr-en: An open-source corpus that contains data crawled from all over the web from the past eight years.
- europarl-v7.fr-en: A corpus that consists of the proceedings of the European Parliament from 1996 to present, all of which are published online. This corpus was originally created for research in SMT (Koehn, 2005).
- giga-fren.release2.fixed: A parallel corpus created for the Workshop on Machine Translation (WMT) 2010 that consists of data crawled from Canadian, European, and international websites.

- news-commentary-v12.fr-en: A parallel corpus of news commentaries provided by the WMT, originally for training SMT systems.

Information about the size of these datasets is shown in Table 1.

Filename	# Lines	# Words	# Characters
commoncrawl.fr-en.en	3,244,152	70,730,355	434,655,470
commoncrawl.fr-en.fr	3,244,152	76,690,762	500,374,763
europarl-v7.fr-en.en	2,007,723	50,263,003	301,523,301
europarl-v7.fr-en.fr	2,007,723	52,525,000	346,919,801
giga-fren.release2.fixed.en	22,520,376	575,753,731	3,789,873,031
giga-fren.release2.fixed.fr	22,520,376	672,168,058	4,565,271,815
news-commentary-v12.fr-en.en	258,432	5,711,996	36,645,477
news-commentary-v12.fr-en.fr	258,432	6,722,644	45,439,337
Combined training corpora	**56,061,366**	**1,510,565,549**	**10,020,702,995**

Table 1 – Size of training corpora

2.2.2. Data preparation

Pre-processing, also known as "data preparation," consists of all the steps required prior to training our systems. It includes cleaning up the corpora (2.2.2.1), segmenting the texts into subwords to allow for an open vocabulary (i.e., a vocabulary made up of words that the systems had not previously seen in the training data) (2.2.2.2), embedding the text (2.2.2.3), connecting newly obtained subwords to form a new text, applying a pre-determined number of merge operations to reorder words within a sentence, and, lastly, applying the subword pre-processing model to our corpus.

All these steps have a direct impact on the systems and are required to achieve better-trained models (Rikters, 2018). Since the object of this thesis is to compare NMT architectures and not pre-processing methods, we used the same pre-processing methods for all of our models, to ensure that we are evaluating all the architectures from the same baseline.

2.2.2.1. Normalization

To work with text-based datasets, some normalization (or data cleaning) of character encodings is required. Character encodings refer to the sets of rules that map binary byte strings (made up of 1s and 0s) to the characters or to the text that they represent. For

English to French translation, we used two cleaning scripts. The first script is an in-house script developed for Portage (clean-utf8-text.pl) that, as its name suggests, cleans UTF-8 (currently the standard text encoding). This Portage script also performs additional whitespace, hyphen, and control character normalization.

The second script is Moses's remove-non-printing-char.perl and it was originally written for the Moses SMT system.[18] It complements Portage's script in cleaning UTF-8.

2.2.2.2. Tokenization

Tokenization (or word segmentation) consists of dividing the texts into sub-units (called tokens) that are meaningful to the machine. This means that a token will not always be a word, a prefix, a suffix, a root, etc. In fact, since tokenization is a process that is frequency-based and not linguistically motivated, the same tokenization methods can be applied to multiple languages and translation directions.

Tokenization is a very important step in the data preparation process, since it can have a direct effect on BLEU scores, therefore influencing how a system is thought to perform (Post, 2018, p. 2). There are three main approaches to tokenization: naïve tokenization algorithms, rule-based tokenizers, and subword tokenizers. Naïve tokenization consists of splitting a string into tokens on whitespaces and punctuation, a method that is no longer sufficient for today's NMT systems. Rule-based tokenizers are more sophisticated, as they look at smaller sub-units, such as prefixes, suffixes, and infixes, and can be customized (with grammar rules) to reflect the specifics of a given language. Their main weakness, however, lies in their ability to handle rare words. Rule-based tokenizers are not always able to segment words into meaningful sub-units if they have not seen those words frequently. Subword tokenizers have attempted to solve this problem by looking at frequent words and less frequent words differently. Frequent words are treated as a unit, while rare words are identified and divided into sub-units that are hypothesized to best carry their meaning. As of the writing of this thesis, there are four main subword tokenization algorithms: WordPiece (also known as sub-word units),

[18] The script can be found at the following link: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/remove-non-printing-char.perl. For more information about the Moses SMT system, refer to http://www.statmt.org/moses/?n=Development.GetStarted

SentencePiece, unigram language model, and byte-pair encoding (BPE; also known as bigram).

By segmenting texts into subwords, we allow room for an open vocabulary. NMT models prior to 2016 typically operated on a fixed vocabulary and had difficulty translating rare or out-of-vocabulary words[19] (Sennrich *et al*., 2016). To address this weakness, Sennrich, Haddow, and Birch (2016, p. 1715) proposed a new approach that made their "NMT model capable of open-vocabulary translation by encoding rare and unknown words as sequences of subword units." With this approach, systems can create new words that they may not have seen in the training data, based on their knowledge of known subword units such as morphemes or phonemes.

In this project, we carried out two phases of tokenization. In the first phase, we used the rule-based Moses tokenizer (https://github.com/moses-smt/mosesdecoder) to separate punctuation from words. In the second phase, we used SentencePiece (https://github.com/google/sentencepiece) to divide words into subwords.[20]

2.2.2.3. Word embedding

Word embedding, or vectorization, essentially allows us to establish the relationships between the source language tokens and the target language tokens (in numerical form). This step is necessary because, as mentioned previously, neural networks can only process numbers.

During this stage, the sequence of tokens in the input is converted into a sequence of vector representations. To help visualize this, Forcada compared the process of using vectors to represent words to where furniture is placed in a room. With the southwest corner of the room as the origin, he describes each object's position as being represented with a three-dimensional vector consisting of information regarding how far north, how far east, and how far above the ground the object is (Forcada, 2017, p. 295). Some objects will be placed closer together than others and, similarly, some tokens in NMT will be

[19] Out-of-vocabulary words consist of words that are "unknown" (Luong *et al.*, 2015b) to the system, likely because they did not appear in the training data.

[20] While SentencePiece does not require a two-phase tokenization process, we decided to go forward with that process as we got better results in our tests and as per the NRC's recommendation.

more similar than others. Word embedding allows us to establish the relationships between the different tokens found in our input.

In our case, the embedding depended on the model and therefore varied between the architectures we tested.

2.2.3. Our systems

We worked with the Digital Technologies Research Centre of the NRC on training different NMT systems for testing and evaluation. We used Amazon Web Services (AWS)'s Sockeye framework (https://github.com/awslabs/sockeye) to build all of our NMT systems, as the toolkit allows for training of CNN-based models, RNN-based models, and attention-based models (Hieber *et al*., 2018).

Table 2 details some of our configurations.

	CNN	RNN	Attention
# tokens – source		500,435,168	
# tokens – target		697,264,273	
size vocab. – source	30,073	30,004	30,522
size vocab. – target	30,063	30,004	30,522
max. length for source seq. length	61	61	61
max. length for target seq. length	61	61	61
# layers	8	4	6
# attention heads[21]			8
BLEU score (when training stopped)	27.43	30.11	30.03

Table 2 – Information about our NMT systems' configurations

All systems were trained using four NVIDIA V100 GPUs and took a little less than a day to train, with our attention-based system taking the longest, followed by our RNN-based system, and lastly by our CNN-based system. Training stopped once all three neural systems reached comparable BLEU scores (we used Portage's BLEU score, 30.55, as our baseline).

[21] In Doshi's explanation of the Transformer in "plain English", he described attention heads as follows: "In the Transformer, the Attention module repeats its computations multiple times in parallel. Each of these is called an Attention Head." (Doshi, 2021)

Once the systems were fully trained, they were ready to use on the challenge set as examples of their various architectures. The NRC helped us run all of our data through the systems and provided us with the output for analysis.

In the next section, we discuss the limitations of using a challenge set approach to evaluate MT systems.

2.2.4. Limitations of the methodology

Because of our restricted framework and timeline, we will only be looking at one language pair, and will only test one direction (English to French). We chose this language pair and this direction not only because it corresponds to the ones in *La traduction raisonnée*, but also because it accurately reflects the reality of the Canadian translation industry. In 2017-2018, it was estimated that 89.2% of the translation work done at the TB was from English to French (Olivier, M., personal communication, March 21, 2019). As the demand for translation to French is higher, it is only logical that we focus on this direction. We will also only be looking at the translation difficulties on a sentence level, to remain consistent with the current functioning of systems and will be using general language, avoiding domain-specific challenges and examples.

While our approach should allow us to identify some of a system's strengths and weaknesses, it should be noted that it also prevents us from evaluating the overall fluency of a sentence. Indeed, by narrowing our focus, we can look for accuracy in a specific part of the sentence we are evaluating, but have to ignore the potential errors in other parts of the sentence. In other words, we might have to sacrifice evaluation of the overall style and clarity for the sake of judging accuracy, and only accuracy of a specific item, at that. We thus must admit to falling far short of the three elements of evaluation cited by Hutchins and Somers (1992).

Testing only one language pair in one direction is another limitation that we have to recognize, as we are trying to determine whether a system's architecture will influence its performance, but are only testing the system with language-specific challenges. Furthermore, we have to face the fact that, since it is impossible for us to review all of our training data, the errors that we might find could come from the data itself rather than the system's architecture.

Nonetheless, by using this error-focused, bottom-up aproach, we are hoping to identify trends that could highlight how a system's architecture influences its performance and provide evidence to help both developers and end-users. In the next chapter, we discuss how the challenge set was designed and readied for testing.

Chapter 3: Challenge set

After presenting the general methodologies for using challenge sets for MT evaluation and the general characteristics of these sets in Section 1.1.2, in this chapter, we will discuss how we created our own challenge set, first by establishing its structure, then by developing an evaluation method, and finally, by describing each of the translation difficulties that we identified as likely to be tackled by current NMT systems.

Following Isabelle, Cherry, & Foster's linguistics-based challenge set approach to evaluate MT (2017), we created a new challenge set consisting of syntactic and lexical difficulties to evaluate an NMT system's strengths and weaknesses in translating isolated sentences from English to French. While Isabelle, Cherry, & Foster based their challenge set on theoretical linguistics (adopting a top-down approach), we have decided, considering our background and work experience in translation, to adopt a bottom-up approach, where we focused on (human or machine) translation difficulties identified as common in literature intended for translators and writers, or on translation errors that we know less experienced translators often make. These included both syntactic and lexical difficulties (while Isabelle, Cherry, & Foster concentrated on syntactic difficulties), and are consistent with the perspective of professional translators' use of MT in the workplace as well, focusing on the need to make corrections to output to make it usable for publication.

To establish an initial set of translation challenges widely recognized in the field on which to base our challenge set, we complemented our own experience with literature from translation pedagogy, writing for translation, translatability, and computers and language.

3.1. Challenge Set Structure

In this section, we will detail what motivated our choice of challenges, as well as describe how we decided to build our challenge set. A copy of our complete challenge set is included in Appendices A and B.

3.1.1. Selecting challenges

We began with examples from the objectives in *La traduction raisonnée, 3e édition*, and considered those objectives to determine which represented clear translation challenges that we deemed fit and fair for a machine to tackle. To do so, we filtered and eliminated objectives that:

- were difficult or impossible to formalize for a challenge set format (e.g., *Objectif 45: Mots français dans le texte de départ*; *Objectif 75: Textes mal écrits*);
- focused on an individual lexical item that did not form part of a larger class of similar items. These were eliminated because challenge sets are typically not designed to deal with single items, but rather with more generalized phenomena (e.g., *Objectif 32: To control*; *Objectif 33: Corporate*, *Objectif 34: Development, to develop*);
- were not expected to reliably produce easily and objectively distinguishable correct or incorrect answers (e.g., *Objectif 61: Voix passive* states that using the passive voice is sometimes correct in French, depending on the nature of the document or the agent on which the author wants to put the emphasis; *Objectif 62: Tournures nominales, tournures verbales* says that French *tends* to favour nouns over verbs, but Delisle & Fiola warn the reader that this is an assumption);
- required adaptations considered to go beyond the scope of what could reasonably be expected of a current MT system (e.g., that require extralinguistic information to resolve the challenge; most of Partie IX: Difficultés d'ordre stylistique).

After examining phenomena from *La traduction raisonnée* that we recognized as difficult for human translators but potentially useful for a challenge-set approach to MT testing, we then examined problems widely recognized as difficult for machine translation and that also satisfied the conditions outlined above for inclusion in our challenge set (i.e., representing a class of phenomena that were expected to be within the scope of problems potentially resolvable by NMT systems and tested using a challenge

set, and for which correct and incorrect translations could be clearly distinguished). Arnold (2003), while dated in the MT approaches to which it was originally applied, nevertheless identifies some phenomena that are still challenging, including anaphora and scope problems. Koehn & Knowles (2017) found, through a challenge set approach, that NMT systems tend to perform poorly on very long sentences[22] and that they generally have lower quality out-of-domain (i.e., when used for texts or vocabulary from domains other than those they were trained for). L'Homme (2008), while she mainly discusses NLP in general, touches upon difficulties that are transferable to NMT, such as structural ambiguities. Matusov (2019), although his research focused on the English to Russian and German to English language pairs, identified repetitions (i.e., the use of homographs) as a category of translation error, a category that can be found in the English to French (e.g., translating *wedding and marriage* as *mariage et mariage*) or French to English (e.g., *rivière et fleuve* as *river and river*) language pairs as well.

These resources allowed us to identify a number of problematic phenomena, as shown in Figure 6.

They also introduced a number of factors which are not in and of themselves representable as challenges *per se*, but which can be used as a complement to challenges selected and thus to target some issues that are relevant both for translatability (as per the literature on writing for translation) and for the literature comparing various approaches to NMT architectures, such as long-range dependencies (Tang *et al.*, 2018). As part of this set of issues, we chose to create short and long variants of sentences to study the range of dependencies. Sentence length has been proven to affect an NMT system's performance throughout the years (Koehn & Knowles, 2017; Pouget-Abadie *et al*., 2014). Researchers typically observe a drop in quality the longer and more complex a sentence is (e.g., Bowker & Buitrago Ciro, 2015), so we wanted to see if sentence length could specifically affect how accurately our targeted difficulties are being translated.

As part of our research, we defined "short sentences" as sentences containing 15 words or less, and "long sentences" as sentences containing 25 words or more (in our source language—English). These numbers correspond approximately to what

[22] We define what constitutes a long or short sentence later in this section.

researchers have previously used as benchmarks when evaluating MT quality according to sentence length (Pouget-Abadie *et al*. considered a sentence to be short if it has ≤ 20 words and long if it has ≥ 20 words), and are also accurate representations of the average length for an English sentence. (Fan [2007] studied two British corpora, the Lancaster-Oslo/Bergen (LOB) and the British National Corpus (BNC), and found that the mean average length of sentences was 21.1663 words for the LOB and 19.6829 and 19.4486 words for the two sets of sample texts comprised in the BNC).

To study the potential effects of structural complexity observed in the literature, we also tested different sentence structures for some challenges (e.g., anaphora), testing some sentences that included potentially "confusing" items between the parts of the sentences that should agree (e.g., between a pronoun its antecedent), and some that did not. (We refer to these variants as options *with* and *without interruptions*.) By testing with these interruptions, we hoped to determine whether the distance between the dependent item and the antecedent, and the presence of other potential antecedents between the two, plays a predictable role in whether the system correctly interprets the sentence and successfully handles the challenge.

Finally, when our challenges could appear in various positions in the sentence structures, we included sentences in our challenge set for each position, to ensure that the full range of possibilities was explored in case the position affected performance for one or more systems.

Furthermore, as NMT is widely recognized for working with a longer window than SMT and being able to use up to a full sentence (if not more) to assist with disambiguation (Agrawal *et al*., 2018), we also included challenge set sentences which contained items that we considered likely to assist with resolving of some of the ambiguities (e.g., *I left my chocolate in the car and it melted*. [the chocolate is more likely than the car to melt]; *I left my sunglasses case in the car and it melted.* [neither the sunglasses case nor the car would be expected to melt]). In combination with the structural variations described above, this ultimately led to the creation of challenge set variants that were long and short, interrupted and uninterrupted, and with different types of semantic cues for a number of the relevant challenges.

In Figure 6, we present a breakdown of our challenge categories. We have included some, but not all, other possible translation problems to help illustrate that the challenges retained are only some among many possibilities.

3.1.2. Building the challenge set

We tested three meanings for each of our polysemous words and therefore counted 18 challenges in total: 12 related to ambiguity in the source language, one related to homographs, two related to scope issues, and three related to anaphora.

Challenge category	Challenge type	Challenge item	Challenges
Lexical	Ambiguity	*as*	S1-3[23]
		while	S4-6
		when	S7-9
		with	S10-12
	Homographs		S13
Syntactic	Scope	modifier	S14
		conjunction	S15
	Anaphora	*it*	S16
		they	S17
		these	S18

Table 3 – Distribution of the challenges

For each of our 18 challenges, we created eight sentences: four short and four long, according to the sentence length definitions that we set out in the previous section. Within these short and long variants, we created other subcategories that varied depending on the challenge. In some challenges we tested different placements of the challenging element in the sentence; in others, we tested different locations of the cue determining the gender and/or number agreement in the sentence. These variants will be further detailed in their respective sections describing the specific challenges.

[23] For more detail on the numbering of challenges in the challenge set, see section 3.3.1

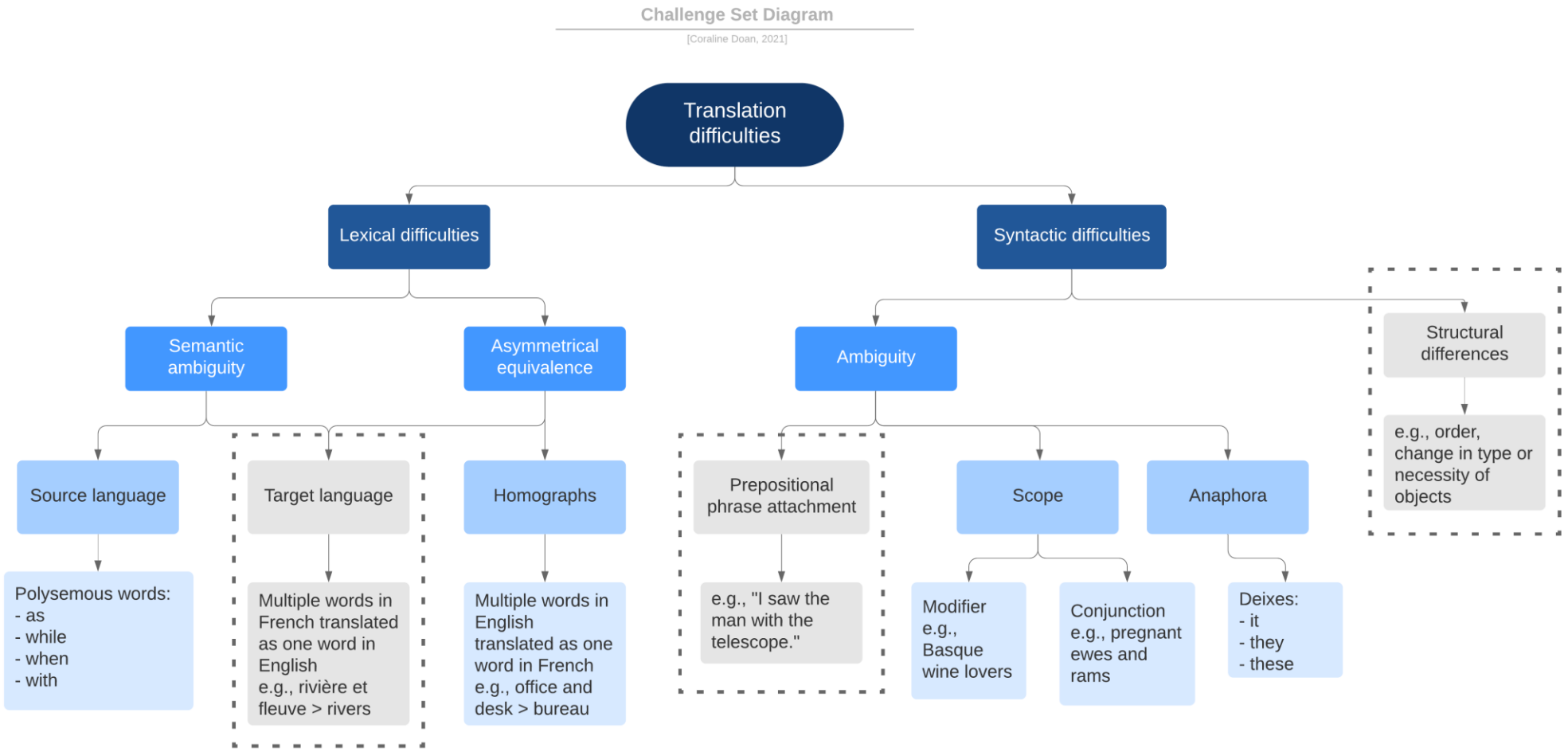

Figure 6 – Overview of the challenge set categories

We also excluded vocabulary that is domain-specific because mistranslation of rare or highly specialized words is not indicative of a specific NMT system's performance, but is rather characteristic of all NMT systems in general (Luong *et al.*, 2015b). Furthermore, this characteristic does not stem from the system's architecture, but rather from the type and amount of training data that is being fed into the system. NMT systems tend to be trained using small vocabularies, which makes it unlikely that they have even come across these highly specialized words (more than one or twice at most) unless the training data was selected specifically to include them. Because of how NMT systems analyze input, unfamiliar words can not only lead to mistranslations of the items themselves, but also affect the processing of the rest of the sentence and create problems elsewhere. It should also be noted that domain-specific-related errors often stem from individual items. Luong and his team have demonstrated that using a simple alignment-based technique where they would link their MT system to a dictionary in a post-processing step could solve the problem, highlighting once again that translation of highly specialized words would not be a relevant category in a challenge set.

All in all, in designing our challenge set, we tried to ensure that the sentences would (insofar as possible) produce clearly correct or incorrect outputs to allow for unambiguous evaluation.

In the following sections, we describe our evaluation method, as well as the lexical and syntactic difficulties that we have identified for our challenge set. We explain how these expressions represent a translation difficulty, how they might be inaccurately translated, then provide a better alternative along with a justification.

3.1.3. Difficulties excluded from the challenge set

The translation difficulties that we have decided to include in our challenge set are only a selection of the difficulties that have been raised in literature.[24] In terms of syntactic difficulties, structural differences between the languages are another translation difficulty category that is often mentioned by authors (Arnold, 2003, p. 119), but that we decided to exclude from our challenge set. Structural differences include a variety of challenges involving differences between typical SL and TL sentence structures at the level of features such as word order, types of objects associated with verbs, and optional or required expression of arguments of verbs. The classic example of word order differences is the shift from an adjective preceding a noun in English (e.g., *the usual order*) to following the noun in French (e.g., *l'ordre typique*). A difference in the type of objects associated with verbs can be seen in the transformation of *I am listening to the music* to *J'écoute la musique*. A difference in the expression of verb arguments can be seen in the compulsory presence of the object *us* in a sentence such as *This data allows us to draw conclusions,* whereas the corresponding object is optional in a sentence such as *Ces données (nous) permettent de tirer des conclusions*.[25] These challenges may not necessarily go beyond an NMT system's translation capacity (the transformation of adjective + noun compounds is generally very well handled by MT, for example). When they do, however, they may be particularly hard to evaluate objectively, as style often has to be considered. For example, when we look at word order, *He gave his girlfriend an orange* could be translated as *Il a donné à sa copine une orange* or as *Il a donné une orange à sa copine*. Both translations are technically correct, but the latter is more idiomatic as the direct object precedes the indirect object in French. Given the challenges in evaluating the more recurrent issues such as the order of objects, and the relative rarity

[24] For example, although we identified asymmetrical equivalence (term used specifically in our challenge set, not an established term in the field) as being a category, we recognize that the difficulties included in our challenge set do not reflect all the translation problems that one might encounter in English and French translation. Furthermore, while we have noticed numerous cases where NMT system would omit words, we have found few examples where it would add words, let alone suggest other possible translations. For this reason, we have decided to exclude (potential) asymmetrical equivalence in the target language as a category from our challenge set (e.g., we believe the chances of an NMT system translating *river* as *rivière et fleuve* are low).

[25] A similar structure in English, such as *This data allows to draw conclusions*, would be considered incorrect.

of other types of structural differences, we ultimately decided to omit structural differences from our challenge set.

Finally, we have also decided to not include ambiguity stemming from prepositional phrase attachment since a literal translation, though still ambiguous, could still be considered correct. For example, in the sentence *I saw the man with the telescope*, the reader can wonder whether the subject was observing the man with the aid of a telescope, or whether the subject was observing a man who was carrying a telescope. In both cases, a literal translation such as *J'ai vu l'homme avec le télescope*, which still renders the ambiguous aspect found in the source, would be correct. In order to produce a translation void of ambiguity, we would need extra-sentential information, which is not within the scope of our study. Therefore, as we are unable to evaluate the translation to such sentences, we have decided to exclude them from our challenge set.

3.2. Evaluation Method

As discussed in 2.2.4, to evaluate our challenge set, we focused only on the particular item that we were targeting for the challenge. As a preliminary step, we assessed our sentences with Google and DeepL, two free online machine translation services, to ensure it would be relevant to include them in our challenge set. Although these two systems are not the focus of the trial, they helped to ensure that we would likely be able to distinguish correct and incorrect translations, and allowed us to foresee potential difficulties in evaluation. For example, we left out some of the translation difficulties that the current systems could repeatedly handle without issues (e.g., *as* introducing an apposition: *As a botanist, I wouldn't recommend watering those plants every day*).

Based on the literature and observations of the two systems, as well as our own experience, we identified a likely but incorrect translation to highlight the potential for error; then, we assigned to each sentence a reference (correct) human translation. It should be noted that this reference translation only serves as an example of a correct translation and is not the only solution considered correct, nor the one specifically expected of the systems.

Once these preparatory steps were completed, the challenge set was submitted to each of our experimental systems, and one more time to Google and DeepL.[26]

In the results, each translation produced by an MT system was assigned one of two levels:

- Correct (✓) (i.e., corresponding to the reference translation, or judged equally acceptable, i.e., not requiring postediting of the specific challenge item to achieve publishable quality);
- Incorrect (✗) (i.e., corresponding to the expected error, or judged equally unacceptable, i.e., requiring postediting of the specific challenge item to achieve publishable quality);

We also marked some of our obtained translations with an asterisk (✓* or ✗*): these are cases where the translations either did or did not require post-editing of the translation of the challenging element (and thus were classified as incorrect or correct, respectively), but did not allow us to judge precisely how the systems handled the targeted challenge. These mostly occurred when a system produced a solution that we were not expecting, but that was acceptable nonetheless, or when a system omitted the challenging element entirely.

3.3. Lexical difficulties

Semantic ambiguity, while widely acknowledged as challenging for MT, is not always a natural choice for challenge-set-based evaluations. As challenge sets aim to evaluate MT's handling of recurrent phenomena, rather than of specific items (which may often be adjusted simply by adding conditional statements),[27] as discussed in Section 3.1.1, many semantic ambiguities would fall beyond the scope of such an approach. We argue,

[26] We ran our sentences through Google and DeepL a second time once all the sentences had been created to reflect the state of the systems on the same day (August 11, 2020), and to account for any changes since the previous tests.

[27] Conditional statements are simple programming commands for handling decisions that can be made consistently according to a simple rule. Here, we are talking about using a simple "if-else" construct for translation of specific items. For example, if we notice that a system translates *identify a solution* as *identifier une solution* (a translation deemed inaccurate according to Objectif 35 of *La traduction raisonnée*), we could add to an NMT system a conditional statement for that phrase to be consistently translated as *trouver une solution* instead.

however, that one exception to this is grammatical words, which belong to a relatively closed set that shares many key characteristics, are highly frequent and widespread, and are recognized as highly polysemous (Sumita & Iida, 1992, p. 86) and thus highly problematic for MT. Moreover, their senses are recognized as being closely linked to their context (Delisle & Fiola, 2013, p. 435), specifically the other items they link in a sentence. As NMT can purportedly make better use of the broader sentence context to process semantics and choose adequate equivalents for polysemous items than previous MT approaches, including a sample of grammatical words in a challenge set allows us to test the systems' ability to resolve recurrent and difficult problems of semantic ambiguity.

In designing our challenge set, we defined lexical difficulties as difficulties relating to the meaning of a word or expression in a given context. Since we hope to create a challenge set that can offer reproducible data and will stand the test of time (i.e., rather than looking at individual cases that can be solved with conditional statements, concentrating on recurrent and more generalized phenomena), we decided to work with lexical difficulties deriving from semantic ambiguity and asymmetrical equivalence (between languages).

When looking at ambiguity problems, we focused on ambiguity in the source language (SL), since we are developing a unidirectional challenge set (for English to French). When the systems are facing a lexical ambiguity in the SL, we expect them to render the wrong word in the translation should they fail the challenge. For example:

- **Source:** *While* I enjoy your company, I'm going to have to ask you to leave.
- **Correct translation:** *Bien que* j'apprécie votre compagnie, je vais devoir vous demander de partir
- **Incorrect translation:** *Pendant que* j'apprécie votre compagnie, je vais devoir vous demander de partir.

Our selection of semantically ambiguous lexical items for the challenge set was inspired by *La traduction raisonnée*, in which it is noted that grammatical words such as prepositions can often be highly ambiguous, often leading to mistranslations even by human translators. Although, in the book, these words are categorized under "Difficultés

d'ordre syntaxique",[28] we believe that according to the classification we use for our challenge set, they are more appropriately considered under lexical difficulties, because the challenge in translating these words accurately relies on the various meanings they take on in different contexts.We found that words such as *as*, *while*, *when*, and *with* were good candidates for our challenge set as they are commonly found in all types of texts.

As mentioned previously, we selected three relationships and/or aspects best suited for inclusion in our challenge set (according to the criteria set out above), and developed a set of challenge set sentences illustrating each of the senses. We also took care to reflect variations in length and in the placement of these items, which in most cases can appear either at the beginning of the sentence or between the two propositions (or other items) they connect. We thus constructed eight sentences for all of these categories: four short sentences and four long sentences. When possible, we also included variants such as:

- Two short sentences with the challenging element before the connected items;
- Two short sentences with the challenging element between the connected items;
- Two long sentences with the challenging element before the connected items; and
- Two long sentences with the challenging element between the connected items.

3.3.1. *As*

According to Delisle & Fiola (2013, pp. 435-441), while translating *as* as *alors que* might be a common "default" solution, it is only acceptable in certain contexts, specifically when the idea of simultaneity is being expressed. In other cases, there are more precise, less abstract translations available in French that reflect other nuances of meaning. In selecting senses for testing in the challenge set, we targeted three possible meanings that

[28] Most of these examples may be considered to illustrate both part-of-speech and semantic ambiguities, as the forms correspond to more than one possible part of speech (e.g., conjunction, preposition, adverb) as well as more than one meaning. In line with Delisle's presentation and the focus on translation, however, we will concentrate on the semantic nuances and the resulting requirement for different equivalents, rather than on the grammatical specificities. A similar choice is made in the *LDOCE Online*, which in most of these cases combines multiple parts of speech into a single entry with a shared set of senses.

were different enough for a clear evaluation, and also judged to be fair for a machine to tackle. We looked at cases where *as* expresses either simultaneity (e.g., *I flinched as the cold water hit me*.), a cause (i.e., when it introduces a justification) (e.g., *I locked the door, as I was the last person leaving*.), or a progression (e.g., *As time goes by, my memory fades.*). We left out *as* expressing a state or quality because we found the nuance to be too subtle for a machine most of the time. In fact, some of the examples given for *as* expressing a state or quality could arguably also be interpreted as examples of *as* expressing a cause (e.g., *During the 1980s, as governments cut back, museums turned more and more to the private sector*) (Delisle & Fiola, 2013, p. 438). Furthermore, we considered that idiomatically rendering this meaning requires more adaptation than could reasonably be expected of an MT system. In cases where *as* takes on this meaning, it will not always have a corresponding word or phrase in French, but will instead need to be expressed through other elements in the TL, such as past participles and gerunds (e.g., *Dans les années 1980, les subventions gouvernementales se raréfiant, les musées ont fait appel de plus en plus au secteur privé* [p. 438]), among others. In short, a certain level of linguistic acrobatics is required to accurately render this meaning in French and, according to our knowledge of current systems, such sentence reorganization surpasses the current systems' capacity.

When it expresses simultaneity, *as* can be translated as *alors que*, *quand*, or *au moment où* (e.g., *J'ai sursauté quand l'eau froide m'a atteint*). When it introduces a justification, *as* would normally be translated using a logical connector expressing a cause, such as *puisque*, *parce que*, *car*, etc. (e.g., *J'ai verrouillé la porte parce que j'étais la dernière personne à quitter*). Finally, when it is used to express progression, *as* should be translated as an adverbial or conjunctive phrase such as *au fur et à mesure que* or *à mesure que* (e.g., *À mesure que le temps avance, mes souvenirs s'estompent*). The period of time in which the two actions linked by *as* overlap is not short, therefore, translating *as* as *alors que* in this context would be incorrect. *As* is rather introducing an action that is ongoing or in development.

The challenge set sentences created to test the systems' handling of the ambiguous word *as* are shown below. Throughout the challenge set, the sentences are numbered by challenge (e.g., with *S1* indicating the set of sentences that correspond to

the first challenge), and each individual challenge sentence is labeled with a letter (e.g., with *a* corresponding to the first sentence containing the challenge, *b* to the second, and so on). Short sentences (a–d) are presented before their longer variants (e–f), and additional characteristics of the sentences are indicated by headings where relevant (e.g., in S1a–d, where the placement of the challenge item has been purposely varied to ensure that different possible structures have been tested). The English language (*Source*) sentence is presented first, followed by the human reference translation (*Ref*), with the challenge item and its equivalent indicated in bold in each one.

3.3.2. *While*

According to Delisle & Fiola (2013, p. 443), the conjunction *while* is widely used, and also widely translated as *bien que* or *alors que*, although it does not always express the idea of opposition. They cite temporality, concession, opposition, and explanation as the four meanings *while* can take on, and suggests various translation options for each meaning. Since we are evaluating the potential of NMT, for our challenge set, we had to take into consideration that some nuances might be too subtle for a machine to recognize and convey. As such, we decided to leave out the explanatory meaning as we felt that, judging from the examples given in *La traduction raisonnée* (e.g., *Honest and legal immigrants who are waiting patiently in the queue are penalized, while the smuggles 'refugees' claims are processed.* [p. 444]), the uses were were too ambiguous and too close in meaning to those given for *opposition* (e.g., *Ontario farmers will benefit from better access to the U.S. market, while the interests of the dairy producers are safeguarded.* [p. 444]) to be clearly differentiated even by many humans, and therefore to be used to objectively evaluate NMT performance.

While can be used to convey many aspects of temporality, but it always links two actions occuring simultaneously, whether in the present or in the past (e.g., *I found a lucky penny while walking*). As a result, when it takes on this meaning, *while* would normally be translated as *en*, *pendant que*, *tout en*, etc., or using a gerund to express simultaneity (e.g., *J'ai trouvé un sou chanceux en marchant*). When *while* takes on a concessive meaning (e.g., *While I understand your difficulty, I can't move the deadline*), it introduces an acknowledgement and should be translated as *bien que*, *malgré que*,

même si, *quoique*, or *si* (e.g., *Bien que je comprenne ta difficulté, je ne peux pas changer la date limite*). In cases where *while* expresses an opposition (e.g., *Blue is a cold colour, while red is a warm one*), it would instead be translated as *alors que*, *mais*, *quant à*, or *tandis que* (e.g., *Le bleu est une couleur froide, tandis que le rouge en est une chaude*).

3.3.3. *When*

Although *when* always conveys an aspect of time, there are subtle nuances expressed by the word that make its translation more complex. We will be looking at three uses of *when*: to establishe a causal link, to express continuity, and to convey a meaning similar to "in spite of" or "even though".

Delisle explains that, although *when* is often translated as *quand* or *lorsque*, doing so does not always accurately convey the relation between a cause and its effect. Translating *whe*n as *lorsque* or *quand* implies that two actions occurred simultaneously, which is not always true, as highlighted in his example *He won this medal <u>when</u> he crossed the enemy lines.* In this case, he won the medal *because* he crossed the enemy lines, not precisely at the time when he crossed the enemy lines. Therefore, it should be translated as *Il a mérité cette médaille <u>pour</u> avoir traversé seul les lignes ennemies* (p. 449).

Even in cases where the actions did occur simultaneously or almost simultaneously, Delisle and Fiola argue that the causality is not well established by *lorsque* or *quand*. *Lorsque* and *quand* only carry a temporal aspect—the reader has to infer causality, which sometimes can be ambiguous. For example, if we translate *Two workers were injured when a fork-lift fell over* (p. 449) as *Deux ouvriers ont été blessés lorsqu'un chariot élévateur s'est renversé*, one can logically suppose that the fork-lift fell on the workers and injured them. However, an unacquainted reader could assume that two workers got injured *while* a fork-lift fell. In this case, it is important to accentuate that one action caused the other, i.e., *Deux ouvriers ont été blessés par suite du renversement d'un chariot élévateur*. Taking these points into consideration, he established that adequate translations of the causal *when* would be *par suite de*, *pour*, or *parce que*, or alternatively could be achieved in a sentence structure that conveys causality (e.g., *Le renversement d'un chariot élévateur blesse deux ouvriers*.) (p. 449).

Consequently, to highlight any inaccurate translation, we purposely created sentences where the events could not be logically interpreted as being simultaneous. For example, *She was diagnosed with epilepsy when she collapsed* cannot be translated as *Elle a reçu un diagnostic d'épilepsie lorsqu'elle s'est évanouie*, since it would be impossible for the subject to receive her diagnosis the second she collapses. Instead, a more precise translation would be *Elle a reçu un diagnostic d'épilepsie parce qu'elle s'est évanouie*.

Similarly, when *when* expresses continuity (e.g., *They did not return home till nine o'clock, when they had a light supper*. [p. 450]), it especially cannot be translated as *lorsque* or *quand*. Delisle and Fiola add that, although it is tempting, translating *when* as *alors que* to express two actions occurring one after the other is also incorrect, as *alors que* usually expresses simultaneity and opposition. For these reasons, they suggest translating *when* by coordinating the actions, one after the other (i.e., using *et* or *puis*) (e.g., *Ils ne rentrèrent qu'à neuf heures et prirent alors une collation*) (p. 450). Since *when* is usually used to introduce the second clause in a sequence of actions, it is normally found mid-sentence. For this reason, exceptionally for this translation difficulty, we did not have a variant in which *when* is placed at the beginning of the sentence. For instance, the sentence *They swam until their fingertips were wrinkly, when they got out of the pool* should be translated as *Ils ont nagé jusqu'à ce que le bout de leurs doigts soit fripé, puis sont sortis de la piscine*, because translating it as *quand ils sont sortis de la piscine* would imply that their fingertips became wrinkly once they got out of the pool, which is not the case.

While analyzing our sentences for the meanings cited above, we came across an additional sense, not mentioned in *La traduction raisonnée*, that we found worthy of addition: when *when* takes on the meaning of "in spite of the fact that" (Merriam-Webster Online, n.d.) or "even though something is true" (Longman, n.d.). We decided to include this meaning because it well suited for a challenge set—more so than the other meanings mentioned in the book (i.e., *when* as a relative pronoun or adverb, or following *hardly*, *barely*, *scarcely*). When *when* is used as a relative pronoun or adverb, it is much less likely to be mistranslated by current systems (as assessed with Google Translate and DeepL). When it is following a set list of words, it is not fit for a challenge set because it

is too case-specific. As explained in previous sections, challenge sets are meant to deal with recurrent phenomena and not single items.

When it takes on the meaning of "in spite of the fact that" (e.g., *I can't believe they went to the beach when we've been told to self-isolate*), *when* can be mistranslated; less experienced translators can fail to recognize the meaning and think that *when* is indicating a temporal or causal connection instead. In this situation, however, it would need to be translated as *alors que*, *malgré le fait que*, *en dépit du fait que*, etc. (e.g., *Je n'arrive pas à croire qu'ils sont allés à la plage malgré le fait qu'on nous ait demandé de nous isoler*).

3.3.4. *With*

Delisle and Fiola observed that prepositions are often mistranslated because they show different relationships in different languages (2013, p. 455). They note that *with* is one that is most often incorrectly translated by those who do not master translation techniques (p. 455). While *with* is usually translated as *avec*, and although *avec* can take on either a causal or coordinative meaning, using such a "generic" translation can sometimes create ambiguity. For example, translating *Three Canadians with the American team* as *Trois Canadiens avec l'équipe américaine* could indicate that there are three Canadians in addition to the American team, when what is truly meant is that three Canadians are part of the American team (p. 456). In this example of mistranslation, we realize that the ambiguity in the source was simply transferred to the target—a solution that is often adopted by MT when there is a lack of textual context. For this reason, in designing the challenge set, we had to come up with sentences that contained all the necessary contextual information to clearly identify a specific sense and therefore for MT (in theory) to accurately translate, while still presenting a challenge.

We looked at the same phenomenon as the one presented in the book, but took a different route in our categorizing, opting for semantic divisions instead. In the book, the authors focussed on parts of speech (POS) when they listed the different possible translations for *with*; in an approach more similar to those used for the items above, we considered the semantic implications for translation, and used dictionaries to compensate for the information not found in *La traduction raisonnée* for this item. We identified

three possible meanings that we deemed were fair challenges for a machine: *with* expressing causality (or used to introduce an explanatory clause, according to Delisle and Fiola's categorizing), *with* expressing a particular feeling or physical state (Longman, n.d.), or *with* meaning "in spite of" (Merriam-Webster Online, n.d).

When *with* expresses causality (e.g., *We cancelled our trip to Portugal, with the new travel restrictions now in place*.), it introduces a reason, and thus is usually translated as *à cause de*, *en raison de*, *car*, *parce que*, *comme*, etc., or using a gerund (e.g., *Nous avons annulé notre voyage au Portugal en raison des nouvelles restrictions de voyage maintenant en place*.)

When it expresses a particular feeling or physical state, *with* is normally followed by said feeling or state (e.g., *She was trembling with fear after she heard the loud bang*.) When it takes on this meaning, it cannot be placed at the beginning of the sentence because it needs to be introduced by a verb or adjective (e.g., *shaking with excitement*, *sick with the flu*). Although in some of these cases, translating *with* with *avec* could be acceptable (e.g., *malade avec une grippe*), there are usually other alternatives that sound more idiomatic (e.g., *atteint de la grippe*). In other cases, using *avec* is incorrect and other prepositions should be used instead (e.g., *trembler de joie* and not *trembler avec joie*). Generally, *de* or *par* are alternatives that work in many cases (e.g., *Elle tremblait de peur après avoir entendu la forte détonation*.) We expect current NMT systems to be able to render this meaning because, although the range of items that can follow *with* is still broad enough for the translation to present a challenge, phrases including this meaning can be considered collocations and NMT systems are known for their idiomatic translations.

Finally, when *with* takes on the meaning of "in spite of" (e.g., *With all his flaws, she still loved him*.) it would normally be translated as *malgré*, *en dépit de*, etc. (e.g., *Malgré tous ses défauts, elle l'aimait*.) When it takes on this meaning, it is usually found in sentences with at least two clauses and is used to highlight that the phrase introduced by *with* does not prevent the other phrase from happening.

3.3.5. Homographs

While the difficulties we have discussed in the previous sections involved ambiguity, this is not the only possible lexical issue for MT. Lexical difficulties of certain types can also result in awkward sentence structures that MT may be required to compensate for to produce a usable translation. This can be the case when equivalence is asymmetrical in the SL and TL, i.e., when a single item in one language (e.g., *river* in English; *mariage* in French) corresponds to two or more items in another (e.g., *fleuve* and *rivière* in French; *marriage* and *wedding* in English).[29] When these items occur together in a single SL sentence, the TL results can be rather odd:

- **Source:** People should put more effort into their *marriage* than their *wedding*.
- **Incorrect translation:** Les gens devraient consacrer plus d'effort à leur *mariage* qu'à leur *mariage*.

A human translator facing this problem will usually find a turn of phrase or a synonym to avoid a repetition:

- **Correct translation:** Les gens devraient consacrer plus d'effort à leur *vie de mariés* qu'à leur *cérémonie de mariage*.

Alternatively, other manipulations (e.g., the omission [Yang *et al.*, 2019, p. 6191] of one of the items) may be an option in some contexts:

- **Source:** J'adore faire du camping à côté d'une rivière ou d'un fleuve.
- **Incorrect translation:** I love camping beside a river or a river.
- **Correct translation:** I love camping beside a river.

The machine, on the other hand, will not always realize that there is a repetition and may ignore it, causing the resulting sentence to sound repetitive and unnatural.

In the following sentences, we have identified several English words that are most often translated using a single French word, to test out how current NMT systems handle

[29] These phenomena are known and have been described in the past using varying terminology. For example, Dubuc refers to them as linguistic reflections of different knowledge structures (*découpage de la réalité*) (2002, p. 2, 142), while Arnold talks about cases of semantic ambiguity (different senses of a polysemous lexical unit, or distinct homographs) (2003, p. 118). Ultimately, because we mainly focus on phenomena that are problematic for MT rather than on fine-grained linguistic analysis of the SL, we consider together cases that can potentially cause problematic repetition, without differentiating between the specific linguistic phenomena underlying each case.

the potential repetitions within a sentence. Although this challenge is word-based, our goal is not to analyze the NMT systems' vocabulary, but rather their (apparent) capacity to detect the presence of an unwanted repetition and to find a workaround to avoid it.

3.4. Syntactic difficulties

We defined syntactic difficulties as difficulties stemming from the relationships between words in a given sentence, i.e., arising from a sentence structure and its analysis.

When we looked at structural ambiguity, we considered different types of attachments that could be hard to translate. We focused on sentences that present a surface ambiguity, but can be understood by a human using reasoning or general knowledge (based on cues included in the sentences); these cases are generally acknowledged in the literature to be difficult for MT systems to resolve (Arnold, 2003), but the development of NMT and its capacity for picking up on contextual semantic cues makes it an interesting problem to explore in this type of evaluation. Using sentences that include such cues in the challenge set not only offers the potential for the MT system to use context to disambiguate, but also allows us to judge the acceptability of the proposed translation with certainty.

We identified a number of sub-categories of structural ambiguities. The first two are prepositional phrase attachment (e.g., *I saw the man with the telescope*: is the subject viewing the man through a telescope or is the man carrying a telescope?), and anaphora (e.g., *She bought a muffin this morning, but it was too dry*: does *it* refer to the muffin or the morning?). A third sub-category can itself be subdivided into scope of conjunction (e.g., *pregnant ewes and rams*: does *pregnant* apply to only the ewes, or also the rams?) and scope of modification (e.g., *Basque wine lovers*: does *Basque* refer to the wine or the wine lovers?).

As discussed previously in 3.1.3, prepositional phrase attachment are not a suitable sub-category for our challenge set since in most situations for this language pair, a literal translation of the preposition would result in an ambiguous, though coherent, translation. Furthermore, sentences that include such ambiguity would most likely require reference to extra-sentential context to resolve said ambiguity, which is beyond the scope

of our research and is generally believed to be beyond the capacity of most current MT systems.

In order to study syntactic difficulties and with all these considerations in mind, we have included challenges for the following sub-categories: scope of modifiers [3.4.1.1], scope of conjunctions [3.4.1.2], and anaphora [3.4.2].

In the following sections, we will describe these syntactic challenges and will elaborate on the evaluation methods we followed for each of these challenges.

3.4.1. Scope

Scope challenges in NMT are usually caused by sentence length and by the system's inability to render a fixed-length vector that includes all the information contained in a long and/or complex sentence (Cho *et al*., 2014a, p. 5). This is mostly observed in encoder-decoder models, as the encoder has to remember all the words leading to the potentially ambiguous part, sometimes sacrificing previous elements to focus on the problematic section. In more recent studies, attention-based models have shown superior results when it comes to handling long sentences (Luong *et al*., 2015a).

We have identified two types of difficulties involving scope that, however, do not arise from sentence length, but rather would involve the system's ability to process contextual semantic cues (partially enabled by attention mechanisms, as mentioned in 1.1.5.1.3). We will be looking at scope of modifiers and scope of conjunctions.

3.4.1.1. Scope of modifiers

To examine the scope of modifiers, we created sentences where a modifier would be followed by two or more nouns, and might be interpreted as applying to only the noun that immediately follows it, or to the multi-word term. For example, the phrase *new puppy kit* might be interpreted as *[new puppy] kit,* i.e., a kit for people with a new puppy, *une trousse pour nouveau chiot*, or as *new [puppy kit]*, a new kit for people with a puppy, *une nouvelle trousse pour chiot*. This can result in two scenarios: either the reader can associate the modifier with the correct noun(s) based on common knowledge and logic, or the reader can determine to which noun(s) the modifier refers with cues included in the sentence.

We purposely used modifiers that could only be associated with one of the two nouns (e.g., *dehydrated dog food* where *dehydrated* should be associated with *food* and not *dog*), or that are far more frequently associated with one of the two nouns (e.g., *purple dog ball* where *purple* is more frequently associated with *ball* than *dog*). We also made sure we used two nouns of different genders to allow us to see if the agreement between noun and modifier is correct (e.g., *dog* is masculine while *food* is feminine in French; we can determine if the translation links the modifier with the correct noun based on whether *dehydrated* has been translated as *déshydraté* or *déshydratée*).

In the latter case, the modifiers can reasonably be associated with both nouns, however, the cues included in the sentence can help to indicate which noun it should be associated with (e.g., *These bad movie actors are wasting their considerable talent playing in such cheesy movies*—in this case, the reader can determine that *bad* applies to the movie and not to the actors because we discuss their *considerable talent* afterward).

In both cases, we constructed short and long variants of our sentences. It should be noted, however, that while in previous challenges, we constructed short and long variants with the intention of testing interruptions, in these sentences, we are simply testing the long sentences as a generic sort of "distractor" for the systems. This does not make the local analysis more misleading or harder; it simply means there is more to analyze at once.

3.4.1.2. Scope of conjunction

Similarly to our approach for scope of modifiers, to examine the scope of conjunctions, we developed sentences that include two elements linked by a conjunction. We added a modifier to one of the elements that would not make sense if associated with the second one (e.g., *pregnant ewes and rams*) and examined if the system would translate accurately (e.g., *les brebis enceintes et les béliers*), or group the two elements and associate the modifier with both of them (e.g., *les brebis et les béliers enceintes*).

3.4.2. Anaphora

A grammatical device used to avoid repetition, anaphora involve using a sort of placeholder (anaphor, often a pronoun) to refer to an item that appears earlier in a

sentence (its antecedent). Pronouns are known to be hard for MT to translate, as their link to the antecedent can sometimes be "lost" due to factors such as sentence length or complexity, structural ambiguity, and/or lack of explicit cues signalling the relationship.

We chose to test this phenomenon using three pronouns in particular: *it*, *they*, and *these*. *It* and *they* are known to be hard for a machine to translate (Guillou & Hardmeier, 2016; Hardmeier & Guillou, 2018) as they can have varying translations depending on the context. To permit unambiguous evaluation of our challenge set results, we focused on two factors that should clearly indicate whether a link between the pronoun and its antecedent has been represented accurately: number and gender. For example, for a sentence such as "I served the pie instead of the cake because my kids prefer it," the translation "J'ai servi la tarte au lieu du gâteau parce que mes enfants la préfèrent," unambiguously indicates a (correct) link between *la* and the feminine antecedent *tarte* rather than the masculine *gâteau*.

For sentences with *it*, we added a layer of difficulty by creating sentences that included interruptions (i.e., for sentences with interruptions, if *it* refers to an antecedent that is feminine, we included a masculine noun between the antecedent and *it*, and vice versa).[30] It is nevertheless expected to be feasible for MT systems to avoid such pitfalls in the proposed sentences. For example, in the sentence *I bought a new mouse to go with my new keyboard, but it is defective as the right click doesn't work*, it is clear to a human that the pronoun *it* refers to the mouse, as a keyboard usually does not feature a right click. It is also presumed that a machine could be capable of linking *right click* to *mouse* rather than *keyboard* through word embedding. *It* should therefore be translated as the feminine *elle* (i.e., *J'ai acheté une nouvelle souris pour aller avec mon nouveau clavier, mais <u>elle</u> est défectueuse car le clic droit ne fonctionne pas*.) However, with *keyboard*, a masculine noun in French, acting as an interruption and placed between *it* and its antecedent *mouse*, there is a chance the machine may confuse the referential elements and translate *it* as a masculine pronoun (e.g., *J'ai acheté une nouvelle souris pour aller avec mon nouveau clavier, mais <u>il</u> est défectueux car le clic droit ne fonctionne pas*.) We particularly expect

[30] The example sentence in the paragraph above is a good example of interruption, as *cake* appears between the pronoun and its antecedent, *pie*. A system that associated the pronoun with the closer noun *cake* (which might be expected in systems that do not successfully complete an analysis of the full sentence) would tend to translate *it* as *le*.

this error to occur in our longer variants, given the increased structural complexity and the greater separation of the pronoun and antecedent.

For sentences with *they*, we ensured that we reflected a variety of possible positions for the cue that helps a reader determine whether *they* refers to a feminine or masculine noun.[31] In some sentences, the cue is found after *they* [cataphora] (e.g., When *they* signed the contract, *the boys* didn't read the fine print), while in others, it is found before [anaphora in the stricter sense] (e.g., *The boys* didn't read the fine print when *they* signed the contract). While we only looked at gender agreement in our sentences with *it*, we also looked at agreement in number in our sentences with *they*. A sentence is only considered correct if the system translated both the gender and the number correctly. To add difficulty to the challenge, we sometimes used nouns that are typically biased and associated with a certain gender (e.g., *nurse* is typically associated with the female gender and translated as *infirmière*), but added a textual cue that would indicate the actual gender for the noun (e.g., the *nurses* announced they were going on *paternity* leave). These cases are extra-difficult examples requiring systems to balance multiple semantic cues that may be contradictory, and may occur at different distances from the pronoun, in order to deal with the challenge and may indicate how a system's architecture influence its performance.

These is a pronoun that expresses plurality, but also refers to an element that is either nearer in location as compared to another item (which may or may not appear in the sentence) or has been mentioned recently (e.g., *Most chocolates are safe, but <u>these</u> contain nuts*). To be considered correct, the translation needs to have an element that renders that proximity (e.g., the use of the adverbial particle *–ci*), as well as be correct in number and gender. This challenge was complicated in some sentences by the presence of interruptions similar to those in the sentences with *it*. For example, in the sentence *Most chocolates are safe and come in sealed boxes, but these contain nuts*, the systems would need to be able to link *these* to *chocolates* and not *boxes* and translate the sentence

[31] *They* in this thesis is tested only as a plural form, although the use of the singular *they* is both increasing and potentially problematic for MT (Guillou & Hardmeier, 2016).

as *La plupart des chocolats sont sécuritaires et viennent dans des boîtes scellées, mais ceux-ci contiennent des noix*.[32]

While of course it is necessarily restricted in size and scope, we believe that the challenge set offers opportunities not only to explore system performance on a number of well-established challenges for machine translation, but also to observe variations in handling of these challenges between systems and in various contexts of MT use (e.g., long and short sentences, various more and less complex structures). Moreover, we hope that these initial results will allow us to identify some strengths and weaknesses of the challenge set approach in general and of this challenge set specifically, which can inform future work. In the next chapter, we will present the initial results of our tests.

[32] An incorrect translation identifying *boxes* as the antecedent of *these* would typically include the feminine form *celles-ci* rather than *ceux-ci*.

Chapter 4: Results and analysis

In this chapter, we provide an in-depth analysis of our results (summarized in Table 4 and reported in Appendices A and B) and an interpretation of the systems' patterns, as we dive deeper into potentially identifying characteristics specific to each of the architectures. We looked at our three NMT systems and used the results from Portage as a baseline representing a hybrid SMT system built with a neural component. We also compared our systems to results from Google and DeepL, extracted on August 11, 2020, as our systems were developed back in early 2019 and there have been significant advances in the field since we last updated our models. In comparing these results, it is also important to recall that the training data for these two systems is different from the corpora used for the other four systems (and far exceeds the corpus size we used).

Table 4 below shows the proportion of challenge set sentences correctly translated by the experimental and reference systems, for the two main categories of challenges. When interpreting the results, it should be kept in mind that these challenges were designed to be problematic for the systems, so a fairly high proportion of failures is not unexpected.

	Hybrid SMT	**CNN**	**RNN**	**Attention**	**Google**	**DeepL**
Lexical (104) (S1-S13)	25 24.0%	23 22.1%	39 37.5%	42 40.4%	54 51.9%	57 54.8%
Syntactic (40) (S14-S18)	18 45.0%	8 20.0%	15 37.5%	8 20.0%	24 60.0%	26 65.0%
Total (144)	43 29.9%	31 21.5%	54 37.5%	50 34.7%	78 54.2%	83 57.6%

Table 4 – Overall number and percentage of correct translations per system

Overall, we found that Google and DeepL performed significantly better than our models, successfully tackling over half of our challenges. Our CNN-based system had the lowest results, followed by our hybrid SMT system, which had accurately translated 29.9% of the sentences. Our RNN-based system and our attention-based system achieved similar results.

We noticed that Google and DeepL were stronger at translating syntactic difficulties than lexical difficulties, while results for these two categories fluctuated more for the other systems. Our hybrid SMT system performed better in the syntactic challenges, while our attention-based system had better results in the lexical categories. Our CNN-based and RNN-based systems performed similarly (or identically in the RNN-based system's case) in the two categories.

In sections 4.1 and 4.2, we provide an in-depth analysis of all of our results for our lexical and syntactic difficulties. These results may be informative for those who are interested in performance for specific items or phenomena, solutions (including unexpected and unusual ones) used by the various systems in dealing with some challenges, and the reasoning behind certain evaluations that may not always be clear at first glance, as well as to those who might consider using the challenge set in future.

In 4.3, we provide an overview of our key findings, broken down in categories such as length of sentences, positions of the challenging element, effect of interruption, etc. These results are more targeted to those interested in the general overview and recurrent observations, and can be useful to users of MT or to those interested in using MT and who are looking to make an informed decision.

4.1. Results – Lexical difficulties

Overall, all the systems did better at translating our examples of semantic ambiguity (polysemy) in the source language than our examples of asymmetrical equivalence with homographs (see Table 5). Our CNN-based system was the weakest overall, while DeepL was the strongest. Among our in-house systems, we noticed that our hybrid SMT and CNN-based systems had comparable results overall, and that the same applied to our RNN-based and attention-based systems, which had better scores than the hybrid SMT baseline.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Semantic ambiguity (96) (S1-S12)	23 24.0%	23 24.0%	38 39.6%	41 42.7%	54 56.3%	53 55.2%
Asymmetrical equivalence (8) (S13)	2 25.0%	0 0.0%	1 12.5%	1 12.5%	0 0.0%	4 50.0%
Total (104)	25 24.0%	23 22.1%	39 37.5%	42 40.4%	54 51.9%	57 54.8%

Table 5 – Number and percentage of correct translations for lexical difficulties

Among the polysemous words, we found that *as* and *while* were clearly easier for systems to translate than *when* and *with*. Errors found in translations involving homographs were mostly repetitions, as we had predicted. In the next sections, we will look at the results in detail and identify the systems' strengths and weaknesses, as well as potential patterns in their performance.

4.1.1. Semantic ambiguity – Source language

We first looked at challenges that consisted of ambiguities in the source language. These challenges included translating the highly and subtly polysemous words *as*, *while*, *when*, and *with*, in various contexts. As shown in Table 6, as was the case in the overall results for this category of challenges, our RNN-based and attention-based systems tended to perform better—sometimes much better—than our hybrid SMT system, as well as our CNN-system. Google and DeepL, however, remained at the top, surpassing our strongest, attention-based system for this category.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
As (24) (S1-S3)	7 29.2%	7 29.2%	20 83.3%	19 79.2%	19 79.2%	23 95.8%
While (24) (S4-S6)	10 41.7%	14 58.3%	14 58.3%	15 62.5%	20 83.3%	15 62.5%
When (24) (S7-S9)	2 8.3%	0 0.0%	0 0.0%	1 4.2%	4 16.7%	5 20.8%
With (24) (S10-S12)	4 16.7%	2 8.3%	4 16.7%	6 25.0%	11 45.8%	10 41.7%
Total (96)	23 24.0%	23 24.0%	38 39.6%	41 42.7%	54 56.3%	53 55.2%

Table 6 – Number and percentage of correct translations per category of polysemous words

It is interesting to note that although *as* and *while* were the polysemous words that our in-house systems managed to accurately translate the most often, results for *as* differed greatly between our hybrid SMT and CNN-based, and our RNN-based and attention-based systems, which, on average, performed much better.

Also worth noting is that *when* was the polysemous word that the systems struggled to translate the most, with our CNN-based and RNN-based systems failing to accurately translate any of the sentences, and with the commercial systems achieving their lowest scores for all polysemous words. *When* expressing causality was mostly translated as a temporal *when*, whereas *when* expressing continuity was translated as a temporal *when* (*lorsque*) or as a *when* expressing opposition (*alors que*).

It would appear that the frequency of senses (and thus equivalents) in use (i.e., senses/equivalents found in the training data) affects the translation performance of certain models, as some systems appear to default to a single sense for some items, while others offer a range of options. Google and DeepL performed significantly better for *with* than the other systems. We think that this is because the meanings we picked for the challenge are less common,[33] and Google and DeepL could have achieved better

[33] *With* expressing causality [S10] is listed fourteenth among the senses identified in the *Longman Dictionary of Contemporary English Online* (n.d.) (hereafter LDOCE), which orders its definitions by frequency of occurrence in its corpus ("About," LDOCE, n.d.)

translations due to their having more training data. This is further supported by the results for *with* in the sense "in spite of" (S12), which is the least frequent of those tested[34] and was inaccurately translated by all systems.

We also think that the complexity of the analysis required to grasp the meaning of a polysemous word directly affects the accuracy of the choice of word. We believe that collocations or local word associations/co-occurrences are more effective solutions to resolve lexical ambiguity than full propositions, as these units are more accessible to NLP techniques (resulting in higher chances of finding readily available solutions in the training data, therefore requiring less memory and less computation). For example, *with* meaning "in spite of" (mistranslated by all systems), requires the systems to take the two phrases linked by *with* as a whole and determine their correlation (e.g., *I still think we should launch the project,* ***with*** *all the risks it entails*.) *With* expressing a particular feeling or physical state, on the other hand, was easier to translate as a local analysis could use close co-occurrents or collocations as markers of this sense (e.g., *consumed with guilt*, *burning with hatred*). The translation was further facilitated by the fact that *with* expressing a particular feeling or physical state is a relatively frequent sense.[35]

In the next sections, we will examine the results for polysemous words in our challenge set and we will try to link the patterns we identify to specific features of each system's architecture.

4.1.1.1. Results for as

When we looked at the overall results for *as* (summarized in Table 7), we noticed first and foremost that our hybrid SMT and CNN-based systems had the lowest scores, while all the other systems translated the majority of the sentences correctly on average.

[34] This was the nineteenth sense in the LDOCE entry for *with*.
[35] This was the fourth sense in the LDOCE entry for *with*.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Simultaneity (8) (S1)	0 0.0%	1 12.5%	7 87.5%	6 75.0%	8 100.0%	8 100.0%
Cause (8) (S2)	7 87.5%	5 62.5%	7 87.5%	6 75.0%	4 50.0%	7 87.5%
Progression (8) (S3)	0 0.0%	1 12.5%	6 75.0%	7 87.5%	7 87.5%	8 100.0%
Total (24)	7 29.2%	7 29.2%	20 83.3%	19 79.2%	19 79.2%	23 95.8%

Table 7 – Number and percentage of correct translations of the polysemous word *as*

DeepL had the highest number of accurately translated sentences: 23 of a total of 24. Next was our RNN-based system, followed by our attention-based system and Google. Furthermore, we were impressed by the fact that, overall, our RNN-based and attention-based systems achieved a score higher than or equal to Google's, especially considering that our systems were developed using older models and less training data.

4.1.1.1.1. *As* expressing simultaneity

For this challenge, we were looking for translations that included solutions such as *alors que*, *quand*, or *au moment où*. Overall, our RNN-based and attention-based systems had no issue translating *as* expressing simultaneity. Our hybrid SMT system as well as our CNN-based system did poorly, with our hybrid SMT consistently translating *as* expressing simultaneity as a causal *as* (i.e., translated as *comme*). Our CNN-based system showcased less consistent behaviour, sometimes translating *as* as the causal *as*, other times ignoring the challenge altogether.

We have summarized our results for short and long variants in tables (such as Table 8 below) where it would be easy to compare how the systems performed in short and long sentences. In these tables, (S) indicates a short sentence, (L) a long one, ✓ indicates a correct translation, ✗indicates a wrong one, and an asterisk (*) indicates cases where the system's solution did not allow us to judge precisely how it handled the targeted challenge.

	"*as*" before the propositions				"*as*" between the propositions			
***As* – Simultaneity**	**1a (S)**	**1e (L)**	**1b (S)**	**1f (L)**	**1c (S)**	**1g (L)**	**1d (S)**	**1h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✓	✗	✗*	✗	✗	✗	✗	✗
RNN	✗	✓	✓	✓	✓	✓	✓	✓
Attention	✓	✓	✓	✓	✗	✗*	✓	✓
Google	✓	✓	✓	✓	✓	✓	✓	✓
DeepL	✓	✓	✓	✓	✓	✓	✓	✓

Table 8 – Results for short and long variants of sentences with *as* expressing simultaneity

Relatively little variation was found in system performance depending on the placement of *as* expressing simultaneity in the sentence; systems generally handled both sets of challenges equally well or poorly. With *as* placed before the propositions in short sentences, there were three occurrences where our hybrid SMT system and our RNN-based system mistranslated *as* expressing simultaneity. In these three cases, *as* was translated as the causal *comme*, which is inaccurate. Translating *as* as expressing causality implies that one element is a cause and the other the effect. However, in S1a for example (*As I took my fries out of the bag, a seagull tried to steal some*), where both the hybrid SMT and RNN-based systems translated *as* as a causal *as*, taking fries out of a bag does not cause seagulls to steal them systematically. When readers read the sentence as a whole, they can deduce that there were most likely specific circumstances (such as the subject being at a location where there are seagulls) that led the seagulls to steal the fries, making this event an event made up of simultaneous actions and not of a cause and effect. For our RNN-based system that translates sentences sequentially, making this link by using cues can be harder, especially when *as* is placed at the very beginning of the sentence and not between the two propositions. In the case of our hybrid SMT system, as hypothesized previously, we believe the system defaulted to translating *as* as *comme*, regardless of the relationships between the propositions, as it also did in S1b.

In S1b, our hybrid SMT system once again translated *as* as *comme*. Our CNN-based system also failed this challenge, but it is worth noting that it failed because it

omitted to translate the challenging element and not because it inaccurately translated said element. It did not use a conjunction to explicitly link the two propositions, making it difficult for us to identify any plausible pattern. Though we know the system did not translate the challenge successfully (i.e., some post-editing would still be required, as determined in 3.2), we cannot hypothesize on the route the system has decided to take, whether the relationship was analyzed as simultaneous, causal, or other.

With *as* placed between the propositions in short sentences, our hybrid SMT system and our CNN-based system did not manage to translate the meaning accurately. Surprisingly, there was also one case where our attention-based system failed to translate this meaning accurately.

In S1d, though we expected our hybrid SMT to fail the challenge, we were surprised to see that this was the only sentence in which our hybrid SMT did not default to translating *as* as *comme*. Instead, *as* was translated as *que*, and we hypothesize that it could be an incomplete translation for *alors que*. If our hybrid SMT system truly failed the challenge because it generated an incomplete translation (and not because it generated the wrong translation), it would mean that in S1d specifically, the system got the correct meaning, but simply failed at translating it. Our CNN-based system failed this challenge in a similar manner, translating *as* as *que*. However, it is worth noting that there is less post-editing required in our CNN translation, as it translated the past continuous by adding *en train* to the sentence. The fact that the system translated the continuous tense shows that it identified an unfinished action (i.e., "as I was restoring an old painting"), interrupted by another action occurring simultaneously (i.e., "I discovered a hidden image"). This strengthens our hypothesis that *que* is an incomplete translation for *alors que* and that the system did get that the meaning expressed by *as* in S1d is simultaneity.

In our long variants, we noted that our hybrid SMT system and our CNN-based system did not manage to translate the challenges accurately. We also noted that our attention-based system failed to translate S1g accurately, as it did for its short variant S1c.

In the long variants, our hybrid SMT system defaulted to *comme* as a translation for all the sentences, including for the long variant of S1d, where the system had translated *as* as *que*. Our CNN-based system also showed results consistent with the short

variants: it translated *as* as *comme* in three sentences (S1e; S1f; S1g), but translated *as* as *que* in the fourth sentence (S1h), as it did in its short variant (S1d).

In S1g, our attention-based system managed to translate the first temporal conjunction included in the sentence (*when* my knees cracked > *lorsque* mes genoux se sont fissurés), but omitted to translate the temporal conjunction *as* (my knees cracked *as* I stood up to go grab my phone > mes genoux se sont fissurés pour prendre mon téléphone). We can hypothesize that the system made this choice due to the fact that multiple temporal conjunctions were included in the sentence, making it harder for the system to build a strong context.

4.1.1.1.2. *As* expressing a cause

As expressing a cause seems to be the meaning that our hybrid SMT system as well as our CNN-based system handled the best out of the various meanings of *as*. For this challenge, we accepted solutions such as *puisque*, *parce que*, *car*, *comme*, etc. We noticed in the last section that our hybrid SMT system consistently translated *as* as *comme*. This makes it hard for us to determine whether the hybrid SMT system has more ease identifying causal links, or simply translated using its "go-to" or default translation for *as* —*comme*, since *comme* coincidentally also expresses a cause.

We noticed more errors in our long variants than in our short variants, and the two sentences that most systems failed to translate accurately were S2a and S2e (short and long variants of the same sentence).

	"*as*" before the propositions				**"*as*" between the propositions**			
***As* – Cause**	**2a (S)**	**2e (L)**	**2b (S)**	**2f (L)**	**2c (S)**	**2g (L)**	**2d (S)**	**2h (L)**
Hybrid	✓	✓	✓	✓	✓	✗	✓	✓
CNN	✗	✓	✓	✓	✓	✗	✓	✗
RNN	✓	✗	✓	✓	✓	✓	✓	✓
Attention	✗	✗	✓	✓	✓	✓	✓	✓
Google	✗	✗	✗	✗	✓	✓	✓	✓
DeepL	✓	✓	✓	✗	✓	✓	✓	✓

Table 9 – Results for short and long variants of sentences with *as* expressing a cause

For the short variants, there were only failed challenges in sentences where *as* expressing a cause was placed before the propositions. All the short sentences with *as* placed between the propositions were translated accurately. As mentioned previously, S2a was the sentence that was mistranslated the most in this challenge, i.e., by three of the six systems (CNN-based, attention-based, and Google). Our CNN-based system and Google both translated *as* expressing a cause as a temporal *as* expressing simultaneity (*Alors que je me suis réveillé* and *Alors que je continuais à être réveillé* respectively). As mentioned previously, this is inaccurate because *alors que* is used to link two simultaneous events that are also opposing,[36] which is not the case in S2a, as the events in the first proposition explain why the events in the second proposition are occurring. By defaulting to *alors que*, the systems are presenting the events as contrasting, which suggests an inaccurate logical relationship and causes a syntactic anglicism.

Our attention-based system, which also failed this challenge, translated *as* as *au fur et à mesure que*. We believe this could be attributed to the fact that *kept being* in the first proposition could also suggest a continous action (as opposed to a repetitive action). The system could have interpreted *as* as expressing progression and translated it as *au fur et à mesure que*.

Google's translation was also wrong in S2b. The system translated *as* as *alors que*, as it did in S2a. We noticed that up to this point, Google has almost always translated *as* as *alors que* (with the exception of S1f, where it translated *as* using a gerund), which makes it difficult for us to determine whether it is its default translation, or if the difference between *as* expressing simultaneity and *as* expressing a cause is too subtle for Google to make the distinction between the two meanings.

All the systems correctly translated *as* expressing a cause in our short variants where *as* is placed between the propositions.

In our long sentences, we found more occurrences of the systems failing to translate the causal *as* than the previous *as* expressing simultaneity. This could be in part attributed to the fact that, according to the Merriam-Webster, "The time-related meaning

[36] According to *Antidote French v4.1*, the conjunction *alors que* is used to link two events, P and Q, occurring simultaneously, with the Q opposing P.

of *as* is more common than the causal meaning of *as*"[37] (Merriam-Webster, n.d.), which could influence the systems' training.

In our long variants with *as* placed before the propositions, we found errors similar to the ones in the short variants (i.e., the systems that failed the challenge all translated the causal *as* as *alors que*). With *as* placed between the propositions, however, we noticed less consistent patterns. In S2g, our hybrid SMT system unexpectedly translated *as* as *as* expressing a state or quality (i.e., *en tant que*) (Delisle & Fiola, 2013, p. 438). In S2h, our CNN-based system produced a translation that is structurally very idiomatic (i.e., *comme suit* followed by a colon to introduce an explanation and/or description), but does not make sense in this sentence. Even though *comme suit* can introduce an explanation, this explanation usually relies on a description of the solution, rather than a description of the reason. In our case, the second proposition, "the path is narrow…", is the reason why "Hikers tend to avoid that trail", not the solution.

All in all, it was interesting for us to note that Google had the worst performance out of the six systems when it comes to *as* expressing a cause, as it seems to default to *alors que* in cases where simultaneity is not necessarily expressed and neither is opposition.

4.1.1.1.3. *As* expressing progression

As expressing progression was mostly mistranslated by our hybrid SMT system and our CNN-based system. The solutions accepted for this challenge included *au fur et à mesure que* and *à mesure que*. Overall, we noticed that, once again, our hybrid SMT system translated *as* as *comme*, with the exception of S3c where it translated *as* as *que*. Coincidentally, our CNN-based system also translated *as* in S3c similarly. S3d and its longer variant S3h were the two sentences that were the most problematic for most systems. In fact, these were the only two sentences that our RNN-based systems did not manage to translate accurately. In terms of performance, our attention-based system had similar results to Google. DeepL had the highest results with eight accurate translations (i.e., 100% of the sentences).

[37] *As* expressing a cause is also the fifth sense in the LDOCE entry for *as*, while *as* expressing simultaneity is the fourth meaning.

At first glance, the systems seem to have performed similarly for the short and long variants.

	"*as*" before the propositions				"*as*" between the propositions			
***As* – Progression**	**3a (S)**	**3e (L)**	**3b (S)**	**3f (L)**	**3c (S)**	**3g (L)**	**3d (S)**	**3h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✗	✗	✓	✗	✗	✗	✗*	✗
RNN	✓	✓	✓	✓	✓	✓	✗	✗
Attention	✓	✓	✓	✓	✓	✓	✗	✓
Google	✓	✓	✓	✓	✓	✗	✓	✓
DeepL	✓	✓	✓	✓	✓	✓	✓	✓

Table 10 – Results for short and long variants of sentences with *as* expressing progression

In our short variants, S3b was the sentence that was correctly translated the most, while S3d was the sentence that the systems struggle to translate accurately the most. The solution that was used by most systems to translate *as* expressing progression in our short variants was *au fur et à mesure que*. There was also one instance where our hybrid SMT system did not default to *comme* as its translation for *as*. In S3c, our hybrid SMT system as well as our CNN-based system translated the progressive *as* as *que*, similarly to how both systems translated the simultaneous *as* as *que* in S1d. While we hypothesized that the systems generated an incomplete translation in S1d (*que* instead of *alors que*), for S3c, we instead hypothesize that the systems generated such a translation because of the incomplete comparison (*comparatif elliptique* [Delisle & Fiola, 2013, p. 417]) *bigger*. We think the systems might have tried to find a second noun to compare the *house* to, but the only other noun present in the sentence was *family*, thus rendering translations as follow: *une maison plus grande que notre famille* and *une plus grande maison que notre famille*. This sentence highlights the inability of our hybrid SMT system and of our CNN-based system to 1- produce a sentence that may contain a less grammatically correct translation by keeping the "comparatif elliptique" (e.g., *une plus grande maison*) or 2- move away from the original sentence structure to produce a translation avoiding the

"comparatif elliptique" (e.g., *une maison plus grande que la précédente*)—although we deem this option unfair to expect from a machine.

Our CNN-based system also resorted to *que* in S3d. However, by looking at how the other systems translated this sentence, we noticed that all of our systems struggled to translate *as* expressing progression in the presence of part-of-speech ambiguity. Though it was not our intention, by using *cooks* as a verb, we accidentally made it complicated for MT systems to interpret the rest of the sentence, potentially creating structural and syntactic ambiguity. All of our systems translated *cooks* as the plural nouns *cuisiniers* or *cuisinières*, which likely caused them to misinterpret the relation between the two propositions. While we intended to express progression, the systems interpreted *as* as expressing a quality or state, with the exception of our CNN-based system, which seems to have, once again, used *que* to compare two nouns (*quantités* and *cuisiniers*). We hypothesize that the one factor contributing to the difficulty in this sentence is that the word *cooks* was probably not found (very often) in the training data as a verb.

In our long sentences with *as* before the propositions, our hybrid SMT system and our CNN-based systems were the only systems that failed to translate the challenge. In S3e, our hybrid SMT system and our CNN-based system both got the meaning of *as* wrong and translated *as* expressing progression as *as* expressing a cause. In S3f, while both hybrid SMT and CNN-based systems mistranslated the sentence, they each translated *as* as expressing a different meaning. As expected from the hybrid SMT system, *as* was translated as a causal *comme*. However, the CNN-based system translated *as* as a simultaneous *as* (i.e., *alors que*). We noticed that in this sentence, our CNN-based system found the positive contraction *we'll* troublesome to translate. It seems as though when this contraction is found in a long sentence with *as* before the proposition, our CNN-based system somehow tries to split the sentence into two (this will be further discussed in 4.4.2). We noticed this pattern in S3e and again here in S3f. In S3e, "require expertise from experienced *workers, and* we'll have to hire" was translated as "besoin de l'expertise de travailleurs expérimentés et nous y *sommes. Il* faut embaucher davantage". In S3f, although the system didn't include a period, it did still include a capital mid-sentence: "As we receive the comments from the people who participated in our workshop last *week, we'll* compile them" was translated as "Alors que nous recevons les

commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous *sommes Les* compiler". We also noticed that in both S3e and S3f, our CNN-based system mistranslated the verb tense of the contraction. The future tense expressed by *we'll* was translated as a present tense in French (*nous sommes*). This could explain the translation for *as* in S3f. If the sentence had been: "As we [action verb in the present tense], we [action verb in the present tense]", *as* could have been expressing simultaneity. Therefore, we can hypothesize that our CNN-system considers that the verbal contraction *we'll* expresses the present tense, which potentially influences its interpretation of the relationship between the two propositions.

In long sentences with *as* expressing progression placed between the propositions, we note that our hybrid SMT system and our CNN-based system were once again outperformed by the other models. Our hybrid SMT system translated *as* as *comme* in both S3g and S3h, like it did in most of the other sentences. Our CNN-based system, on the other hand, displayed a more unexpected behaviour in S3g by translating *as* as *à savoir:*, a structure similar to the one we found in S2h, where *as* had been translated as *comme suit:*. While this structure is very idiomatic in French, its meaning is not appropriate for this sentence as *à savoir:* it loosely translates to *namely* and would generally be followed by a noun or noun phrase, not a clause. In S3h, our CNN-based system translated *as* as *que* like it did in the short variant S3d.

In addition to our hybrid SMT and CNN-based systems failing this challenge, we also found Google to be wrong in S3g, and our RNN-based system to be wrong in S3h. In S3g, Google translated the progressive *as* as a causal *as*. We think this might be because of the comma found before the *as*, which often indicates that *as* means "because". This is an issue that could be further explored in future research, as longer sentences tend to include more commas, which increases the odds of a progressive *as* being placed after a comma and being mistakenly interpreted as a causal *as*. In S3h, our RNN-based system translated the progressive *as* as a simultaneous *as*. While the verb *cooks* was mistranslated as a noun in the short variant, it was accurately translated in this sentence, perhaps due to the larger context provided.

Other observations made in this challenge include the fact that our hybrid SMT system was not able to translate any occurrence of the contraction *we'll* (found in S3a,

S3b, S3e, and S3f). In all four sentences, the system simply reproduced what was in the source, i.e., it kept *we'll* in the French translations (this will be further discussed in 4.4.2).

4.1.1.2. Results for while

For challenges involving *while*, we observed more homogenous results (Table 11), with our hybrid SMT system seemingly performing better than it did with *as*—with the exception of S4, where it struggled to interpret *while* expressing temporality. We also note that our CNN-based system performed better than it did in the previous challenge, successfully translating 14 sentences, as opposed to seven for sentences with *as*. In fact, our CNN-based system generated as many correct translations as our RNN-based system. Our attention-based system also achieved similarly good results as DeepL. Google remained at the top for this challenge, with over 20 correct translations.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Temporality (8) (S4)	0 0.00%	4 50.0%	6 75.0%	5 62.5%	6 75.0%	2 25.0%
Concession (8) (S5)	4 50.0%	7 87.5%	3 37.5%	4 50.0%	6 75.0%	7 87.5%
Opposition (8) (S6)	6 75.0%	3 37.5%	5 62.5%	6 75.0%	8 100.0%	6 75.0%
Total (24)	10 41.7%	14 58.3%	14 58.3%	15 62.5%	20 83.3%	15 62.5%

Table 11 – Number and percentage of correct translations of the polysemous word *while*

There were two sentences (S4g and S5h) that all systems struggled to translate correctly, while there were five sentences (S5b, S5f, S6c, S6d, and S6h) that all systems managed to translate accurately. There is thus variability in performance among all the systems for the two examples of ambiguity in the source language tested so far, which makes it difficult for us, with the data to date, to predict the overall performance of the systems on individual sentences involving this type of challenge.

4.1.1.2.1. *While* expressing temporality

With *while* expressing temporality, we expected to see translations that included solutions such as *en*, *pendant que*, *tout en*, etc. (e.g., *While I was on the phone, the deliveryman rang the doorbell. > Pendant que j'étais au téléphone, le livreur a sonné à la porte.*) Our hybrid SMT system failed to translate this challenge entirely, as it seemed to have defaulted to translating *while* expressing temporality as *alors que* or *tandis que*. We have previously mentioned that these translations are not accurate when the meaning expressed by *while* is temporality. *Alors que* and *tandis que* are only appropriate if an idea of opposition is also being expressed, which is not the case in our sentences. Surprisingly, this was a meaning that DeepL also struggled to translate accurately. In fact, DeepL was second to our hybrid SMT system in terms of poor performance. Our CNN-based system generated four accurate translations, our attention-based system, five, while our RNN-based system tied with Google with six accurate translations.

Most systems (apart from our hybrid SMT system) managed to succesfully translate the challenge in our short sentences, with a few exceptions.

	"*while*" before the propositions				**"*while*" between the propositions**			
***While* – Temporality**	4a (S)	4e (L)	4b (S)	4f (L)	4c (S)	4g (L)	4d (S)	4h (L)
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✓*	✗	✓	✓	✓	✗	✗	✗
RNN	✓	✓	✓	✓	✓	✗	✓	✗
Attention	✓	✓	✓	✗	✓	✗	✓	✗
Google	✓	✓	✓	✗	✓	✗	✓	✓
DeepL	✗	✗	✗	✗	✓	✗	✓	✗

Table 12 – Results for short and long variants of sentences with *while* expressing temporality

In sentences with *while* before the propositions, our hybrid SMT system and DeepL both inaccurately translated the temporal *while* as *alors que* (*Alors que j'étais au téléphone, le livreur a sonné à la porte.*). There was also one case in S4a, where our

CNN-based system ignored the challenge altogether,[38] but nonetheless produced a sentence that we deemed correct enough to not require any post-editing of the challenge item. By using the *imparfait* in the first proposition and the *passé composé* in the second, the system produced a sentence compatible with a sequence of events where the first action is being interrupted by the second, confirming that at one point in time, the two actions were simultaneous.

In short sentences with *while* between the propositions, our CNN-based system also translated *while* as *alors que* in S4d (He fell asleep *while* reading her eighteen-page-long letter > Il s'endormait *alors qu'*elle lisait sa lettre de dix-huit pages). However, in this case, while the translation is semantically inaccurate, the French sentence itself is plausible because of the switch in gender. In the translation, the person reading the letter is a different person than the person falling asleep. The opposition expressed in the French translation could be that *he* fell asleep although *she* stayed awake to read a letter. This shows us that once again, our CNN-based system sometimes produces translations that are more idiomatic than accurate. It was also interesting to note that S4d was the only short sentence in which all the successful systems used a gerund (*en lisant*) to express simultaneity instead of using a separate conjunction (such as *pendant que*).

In our long variants, there were slightly fewer mistranslations in our sentences with *while* before the propositions than in those with *while* between the propositions. However, overall, there were significantly more mistranslations in the long variants than in the short ones.

In the long sentences with *while* before the propositions, all the incorrect translations were because of the use of *alors que*, with the exception of our CNN-based system's translation for S4e, where the system unexpectedly translated *while* expressing temporality as *while* expressing a concession (*While I was cleaning […], I noticed a letter > Même si j'étais en train de nettoyer […], j'ai remarqué une lettre*). This is surprising because *while* expressing a concession is not a meaning as frequently found as

[38] From the source sentence "While I was cleaning my room, I noticed a letter I had never opened," the CNN system produced "J' étais en train de nettoyer ma chambre, j' ai remarqué une lettre que je n' avais jamais ouvert." We consider *en train* to be the translation for *was cleaning* and not *while*.

while expressing temporality[39] and this would be expected to have been reflected in our training data.

There was only one correct translation in our long sentences with *while* between the propositions. In S4g, all the systems mistranslated *while* expressing temporality as *alors que*. In S4h, our hybrid SMT system produced the same inaccurate translation as it did in the short variant S4d by using *tandis que*, while our CNN-based system, albeit still wrong, used a different solution (*tout en la lisant*). In fact, all the other NMT systems that generated incorrect translations used an almost identical solution (*tout en lisant*). Although the use of a gerund to express simultaneity is correct, the addition of *tout en* renders the solution incorrect, as the adverbial phrase conveys a meaning of opposition between the two simultaneous actions.[40] Google was the only system to use *en lisant* alone, making it the only system to generate an accurate translation in S4h.

4.1.1.2.2. *While* expressing a concession

For sentences with *while* expressing a concession, we were expecting solutions such as *bien que*, *malgré que*, *même si*, *quoique*, *si*, etc. The most successful systems for this meaning were DeepL and our CNN-based system, which both produced seven correct translations, out of the eight sentences in total. In fact, this challenge was the one in which our CNN-based system was the most successful, scoring 87.5% in terms of correct translations. Google came in second, followed by our attention-based system and our hybrid SMT system with four correct translations, while our RNN-based system came in last.

All the systems successfully translated S5b and S5f (short and long variants of the same sentence), while all mistranslated S5h. All in all, there was little difference between the systems' performance for the short and long variants.

---

[39] *While* expressing a concession is the fourth sense in the LDOCE entry for *while*, whereas *while* expressing temporality is the first and second senses. The first and second senses express sub-types of temporality in the definition we used. We drew on Delisle to group these two senses together for our purposes.

[40] In Antidote's entry for *tout, adverbe*, it lists the phrase *tout en* as used in *P tout en Q* (followed by a present participle), where P and Q are simultaneous and opposing.

	"*while*" at the beginning of the sentence				"*while*" mid-sentence			
***While* – Concession**	**5a (S)**	**5e (L)**	**5b (S)**	**5f (L)**	**5c (S)**	**5g (L)**	**5d (S)**	**5h (L)**
Hybrid	✓	✓	✓	✓	✗	✗	✗	✗
CNN	✓	✓	✓	✓	✓	✓	✓	✗*
RNN	✗	✗	✓	✓	✗	✓	✗*	✗
Attention	✗	✗	✓	✓	✓	✓	✗*	✗*
Google	✗	✓	✓	✓	✓	✓	✓	✗
DeepL	✓	✓	✓	✓	✓	✓	✓	✗

Table 13 – Results for short and long variants of sentences with while expressing concession

In short sentences with *while* placed at the beginning of the sentence, all the systems that accurately translated this challenge used *bien que* as their solution. In S5a, half of the systems failed to translate the challenge, with our RNN-based system and our attention-based system using *pendant que*, which expresses temporality, and with Google using *tant que*, which expresses a condition (would translate to *as long as*).

In short sentences with *while* mid-sentence, we noticed more irregular and unexpected patterns. In S5c, our hybrid SMT system translated *while* as *tout* (*The actor, while handsome > L'acteur, tout beau*). *Tout* here is used as an adverb and means "entirely", which does not convey the concessive meaning that we were looking for. This is unexpected because *while*, whether it be the noun, conjunction, preposition, or verb, does not have any meaning relatively close to meaning "entirely". Still in S5c, our RNN-based system also mistranslated *while* expressing concession, but as *alors que*, which conveys a meaning of simultaneity and opposition (*The actor, while handsome […], was not a particularly nice person > L'acteur, alors qu'il était beau […], n'était pas particulièrement agréable*).

In S5d, our hybrid SMT system failed to accurately translate the sentence and used *alors que*, which encompasses the contrasting aspect expressed by *while*, but does not encompass the conceding aspect expressed in this context. Our RNN-based and attention-based systems, on the other hand, generated translations that had almost no connection to the source sentence. Aside from the word *hôtel*, the systems' translations

did not include any other element from the source. Our RNN-based system talked about a hotel located near the airport (*L'hôtel est très bien situé, à proximité de l'aéroport*), whereas our attention-based system, which discusses the hotel located near the airport, also mentions four times that it is near the (possibly bus or train) station (*à proximité de la gare, de l'aéroport, de la gare, de la gare et de la gare)*. While we classified these two sentences as incorrect, we were not able to determine whether the system passed or failed the challenge because it was ignored altogether. This is an example of the unexpected and inexplicable solutions we sometimes get with NMT (this will be further discussed in 4.4.2).

Results for our long sentences with *while* at the beginning of the sentence were similar to our results for the short variants, only this time, Google accurately translated the concessive *while* in S5e. As for the systems' performance for long sentences with *while* mid-sentence, it was stronger in S5g than it was in the short variant S5c, but weaker in S5h than it was in the short variant S5d. In S5g, our RNN-based system generated a correct translation using *si*, a solution we had not seen before in this challenge. Our hybrid SMT produced the same incorrect translation by using *tout* as it did in the short variant.

S5h was the only sentence in this challenge that all systems failed to translate accurately. We found the odd patterns identified in the short variants in these cases as well; however, we found them in our CNN-based and our attention-based systems' translations (as opposed to in our RNN-based and our attention-based systems' translations in the short variants). Our CNN-based system correctly translated the first proposition, but ended the sentence after that and started a new sentence where it discusses a hotel located near downtown. Our attention-based system hallucinated, i.e., produced "[a translation] that may be fluent, but completely unrelated to input" (Wang & Sennrich, 2020), and generated the sentence *L'hôtel est très bien situé, très bien situé, à proximité de l'aéroport, de la gare et du centre ville*. In 4.4.2, we will take a closer look at this type of unexpected translation and we will provide our hypothesis as to why the system produced it, including an explanation of why it included a repetition in it.

4.1.1.2.3. *While* expressing an opposition

Possible solutions for this challenge included *alors que*, *mais*, *quant à*, or *tandis que*. Most systems that inaccurately translated the challenge translated *while* expressing an opposition as expressing a concession and used *bien que* or *même si* as solutions.

In terms of performance, Google came in first, having successfully translated all of the eight sentences. Surprisingly, our hybrid SMT system came in second, tied with DeepL, as well as our attention-based system. In third place was our RNN-based system, and in last came our CNN-based system.

For *while* expressing an opposition, sentences where *while* was placed between the propositions, regardless of their length, were correctly translated more often than sentences where *while* was placed before the propositions.

	"*while*" before the propositions				**"*while*" between the propositions**			
***While* – Opposition**	**6a (S)**	**6e (L)**	**6b (S)**	**6f (L)**	**6c (S)**	**6g (L)**	**6d (S)**	**6h (L)**
Hybrid	✗	✗	✓	✓	✓	✓	✓	✓
CNN	✗	✗	✗	✗	✓	✗*	✓	✓
RNN	✗	✓	✗	✗	✓	✓	✓	✓
Attention	✓	✓	✗	✗	✓	✓	✓	✓
Google	✓	✓	✓	✓	✓	✓	✓	✓
DeepL	✗	✓	✗	✓	✓	✓	✓	✓

Table 14 – Results for short and long variants of sentences with *while* expressing opposition

In short sentences with *while* before the propositions, systems that failed the challenge used *bien que* and *même si* as solutions. Since the systems were mostly consistent in their errors (with the exception of our CNN-based system's translation of S6g), we were not able to identify any discernible patterns in the choices of translation of the challenging element itself. However, we noticed other interesting details in the sentences as a whole, which we will be discussing further in 4.4.2.

We had no particular observations for sentences with *while* between the propositions, as the majority of the systems successfully passed the challenge, using either *alors que* or *tandis que*. This differs from our results in our other challenges

involving *while*, as we previously saw that with *while* expressing temporality, there were more correct translations when *while* was placed before the propositions, and with *while* expressing concession, there were more correct translations in the sentences where *while* was placed at the beginning of the sentence.

In long sentences with *while* between the propositions, there was only one inaccurate translation from our CNN-based system in S6g. The system started a new sentence after the first proposition and ignored the challenging element altogether. We considered this translation to be incorrect, as some post-editing would be needed to define the logical relationship between the two propositions. At this point of our analysis, we start to notice that this problem may be recurrent and, therefore, significant for this type of conjunction in this kind of structure.

All in all, *while* expressing an opposition was a challenge that most systems succeeded in translating (in fact, among the meanings of *while* that we selected for the challenge set, this was the meaning that generated the most accurate translations), which makes it hard for us to identify any other clear patterns.

4.1.1.3. Results for **when**

Of all the challenges in our category "Semantic ambiguity – Source language", *when* led to the highest proportion of incorrect translations. In fact, all systems failed to translate *when* expressing causality and *when* expressing continuity. As for the results for *when* meaning "in spite of", all six systems studied only accurately translated 25% of the sentences, combined.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Causality (8) (S7)	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Continuity (8) (S8)	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
"*in spite of*" (8) (S9)	2 25.0%	0 0.0%	0 0.0%	1 12.5%	4 50.0%	5 62.5%
Total (24)	2 8.3%	0 0.0%	0 0.0%	1 4.2%	4 16.7%	5 20.8%

Table 15 – Number and percentage of correct translations of the polysemous word *when*

DeepL produced the largest number of correct translations, with five correct translations in total, which represents only 20.8% of the overall challenge. Google came in second place, followed by our hybrid SMT system and our attention-based system. Our CNN-based and RNN-based systems did not generate any correct translations in this challenge.

4.1.1.3.1. *When* expressing causality

As mentioned previously, none of the systems successfully translated *when* expressing causality. Solutions we were looking for included *par suite de*, *pour*, or *parce que* (e.g., *She was diagnosed with epilepsy when she collapsed and had a seizure. > Elle a reçu un diagnostic d'épilepsie à la suite de s'être évanouie et d'avoir eu des convulsions.*). We also accepted solutions that involved a sentence structure that conveys causality. All the systems used *lorsque*, which is imprecise, as *lorsque* indicates simultaneity, but not necessarily causality.

	"*when*" before the propositions				"*when*" between the propositions			
***When* – Causality**	**7a (S)**	**7e (L)**	**7b (S)**	**7f (L)**	**7c (S)**	**7g (L)**	**7d (S)**	**7h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✗	✗	✗	✗
RNN	✗	✗	✗	✗	✗	✗	✗	✗
Attention	✗	✗	✗	✗	✗	✗	✗	✗
Google	✗	✗	✗	✗	✗	✗	✗	✗
DeepL	✗	✗	✗	✗	✗	✗	✗	✗

Table 16 – Results for short and long variants of sentences with *when* expressing causality

Though we were not able to find patterns in the systems' solutions to the challenge, we were able to identify some recurring errors that seem to be system-specific and that had sometimes already been established in previous challenges. These patterns will be discussed in detail in 4.4.2.

4.1.1.3.2. *When* expressing continuity

For *when* expressing continuity, we were looking for solutions that included coordinating conjunctions such as *et* and *puis* (e.g., *They swam until their fingertips were wrinkly, when they got out of the pool. > Ils ont nagé jusqu'à ce que le bout de leurs doigts soit fripé, puis sont sortis de la piscine.*). Unfortunately for us, none of the systems we tested produced a correct translation. We recognize that *when* expressing continuity is not the word's most common sense—in fact, though there are meanings ranked below it, it is still the third sense in the Longman Dictionary of Contemporary English (LDOCE) entry for *when*. We also recognize that the distinctions between meanings can be quite subtle, even for a human reader. As described in section 3.3.3, we did our best in developing sentences that describe a sequence of events that cannot be simultaneous, but all the systems still failed to translate this challenge accurately, though there was more variety in the solutions than there were in S7. All of the systems alternated between *lorsque*, *alors que*, and *quand*. *Lorsque* and *quand* are unsuitable solutions because they strictly express simultaneity, and *alors que* is inaccurate because it also expresses opposition.

	"*when*" before the propositions				"*when*" between the propositions			
***When* – Continuity**	**8a (S)**	**8e (L)**	**8b (S)**	**8f (L)**	**8c (S)**	**8g (L)**	**8d (S)**	**8h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✗	✗	✗	✗
RNN	✗	✗	✗	✗	✗	✗	✗	✗
Attention	✗	✗	✗	✗	✗	✗	✗	✗
Google	✗	✗	✗	✗	✗	✗	✗	✗
DeepL	✗	✗	✗	✗	✗	✗	✗	✗

Table 17 – Results for short and long variants of sentences with *when* expressing continuity

We found all three solutions in our short sentences with *when* before the propositions. In S8a, all the systems used *lorsque*, but we noticed a slight variation in the verb tense that our CNN-based system used. While all the other systems used the *passé composé*, our CNN-based system used the *imparfait*. This indicates that our CNN-based system interpreted a different meaning than all the other systems. *Lorsque* can take on two different meanings: it can either express simultaneity (i.e., synonymous to *quand*, e.g., *on ferme les yeux lorsqu'on éternue*), or it can express both opposition and simultaneity (i.e., synonymous to *tandis que* or *alors que*, e.g., *la lune apparaît lorsque le soleil se couche*). In this case, our CNN-based system may have interpreted *lorsque* as expressing opposition and simultaneity because the use of the *imparfait* in the first proposition highlights the fact that the action in the first proposition is being interrupted by the action in the second proposition. In contrast, our other systems and Google and DeepL all used the *passé composé* in both propositions, suggesting simultaneity without opposition.

In S8b (*The car will be fixed by 5pm, when I will come to pick it up*), we purposely used a time reference as a temporal cue that might assist with disambiguation. By saying, *The car will be fixed by 5pm*, we imply that the car could be fixed anytime before 5pm, but the guaranteed, ready-to-be-picked-up time is anytime after 5pm. A human can clearly see that the fixing of the car by 5pm precedes the picking up of the car, which means that the actions in the propositions come one after the other and are not

simultaneous. Using *lorsque* or *quand* in this sentence is inaccurate because it implies also picking up the car by 5pm (*La voiture sera réparée d'ici 17 h, quand je viendrai la chercher*). In addition to getting the challenging word wrong, most of the systems also mistranslated *by 5pm*, an element that could have influenced how the systems translated the *when*. Our attention-based system was the only system that successfully translated *by 5pm* as *d'ici 17h*. Our CNN-based and RNN-based systems, as well as Google and DeepL, all translated *by 5pm* as *à 17h*. As for our hybrid SMT, it seems to have struggled with the wording *5pm* because it kept it as is in the target sentence (its translation was *par 5pm*). In sentences where the systems generated *à 17h*, using *lorsque* or *quand* as a translation for *when* is not inaccurate since the temporal cue indicates a specific point in time at which one can come pick up one's car. Consequently, we can argue that in these sentences, the mistranslation of the challenging element was caused by a misinterpretation of the temporal cue. In our attention-based system's case, however, we are not able to explain why the system accurately translated *by 5pm*, but still managed to inaccurately translate *when*.

In our short sentences with *when* between the propositions, we observed no particular phenomenon that could have highlighted why a system mistranslated the challenging element, but we did note some interesting translation choices, unrelated to the challenging element. These will be discussed in section 4.4.1.

The systems produced similar solutions in the long sentences with *when* before the propositions as they did in the short variants.

In S8f, we found the same mistranslation of *by 5pm* in all of our in-house systems, with the exception of our attention-based system. In the long variants, Google and DeepL changed their translation of *by 5pm* to *avant 17h*, but still failed to accurately translate *when*. Translating *by 5pm* as *avant 17h* puts emphasis on the fact that the car will be fixed *before* 5pm and can surely be picked up *after* 5pm. The sequence of events is clearly defined as the first occurring before 5pm, which allows the reader to infer that the second occurs after 5pm. Using *lorsque* or *quand* here is inaccurate as it makes it sound as though the picking up will also occur before 5pm (*The mechanic […] will have the car fixed by 5pm, when I will come to pick it up > Le mécanicien […] fera réparer la voiture avant 17 h, quand je viendrai la chercher*). We were also able to identify some errors in

other parts of the sentence, such as our hybrid SMT system making an agreement error in *la voiture fixé*, an error it did not make in the short variant. We, therefore, note that longer variants might be more problematic for our hybrid SMT system.

In long sentences with *when* between the propositions, the systems mostly all used the same solutions for *when* as they did in the short variants, or used *quand* and *lorsque* interchangeably. It seems as though, for this challenge, the position of *when* does not affect the systems' performance, although this could also be due to the poor results we observed in our short variants making it hard for us to draw a conclusion. We also found a few phenomena in other parts of the sentences that we had identified as being characteristic of some of our systems. These will be discussed further in 4.4.2.

4.1.1.3.3. *When* meaning "in spite of the fact that"

To express the meaning "in spite of the fact that", *when* can only be placed between two propositions (e.g., *I was angry that he left <u>when</u> he said he would stay to help*), and not at the beginning of the sentence (e.g., ***<u>When</u>** he said he would stay to help, I was angry that he left*). Therefore, while the challenge set included the usual eight sentences for this sense, the only variation tested was sentence length. For this sense, we were anticipating solutions that included *alors que*, *malgré le fait que*, and *en dépit du fait que*.

Though our systems did not achieve highly accurate results, our hybrid SMT system and our attention-based system did produce some correct translations, while they did not in the two other meanings of *when* that we tested. DeepL was the most successful system, and was followed by Google. Our hybrid SMT system produced two correct translations, and our attention-based system, only one. Most of the times, the systems failed the challenge because they translated *when* as *quand* or *lorsque*, which only express simultaneity, or opposition and simultaneity, respectively.

In these sentences, we expected the systems to use a prepositional phrase to link the two propositions that would not only express opposition, but also highlight the fact that the actions in one proposition are not being affected by the actions in the other (*J'étais fâché qu'il soit parti, <u>malgré le fait qu'</u>il ait dit qu'il resterait aider*). All the systems that successfully translated the challenging element used *alors que* as their solution.

We, unexpectedly, had more incorrect translations in our short variants than in our long.

When – "in spite of"	9a (S)	9e (L)	9b (S)	9f (L)	9c (S)	9g (L)	9d (S)	9h (L)
Hybrid	✗	✗	✓	✓	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✗*	✗	✗	✗
RNN	✗	✗	✗	✗	✗	✗	✗	✗
Attention	✗	✗	✗	✓	✗	✗	✗	✗
Google	✗	✗	✓	✓	✗	✗	✓	✓
DeepL	✗	✓	✓	✓	✗	✗	✓	✓

Table 18 – Results for short and long variants of sentences with when meaning "in spite of the fact that"

In our short sentences, S9a and S9c were mistranslated by all systems, whereas S9b was accurately translated the most (by three of the six systems we tested). In S9a, all of our systems translated *when* as *lorsque*, while Google and DeepL translated *when* as *quand*. Our attention-based system also failed to translate part of the source sentence and omitted *that he left* in the target sentence, resulting in *J'étais en colère lorsqu'il a dit qu'il restait à aider*. In S9b, though our hybrid SMT system successfully translated the challenging element, we noted that it once again struggled with the verbal contraction *couldn't*, and it left it as is in the target sentence. In S9c (*I was upset she still dated him when I warned her he was a cheater*), all of our systems failed the challenge, but it was interesting to see that our CNN-based system completely ignored the challenging element and reworded the sentence, shifting the meaning of the source and producing a new sentence (*J'ai été bouleversée qu'elle l'avait toujours avertie qu'il était un tricheur*).

In our long sentences, S9g was the only sentence that all systems failed to translate correctly. In S9e, DeepL was the only system to correctly translate the challenge with *alors que*. We also noticed that our CNN-based system failed to translate a section of the target in this sentence (*he knew we needed* was translated as *qu'il savait*). As mentioned previously, S9f was the sentence that generated the largest number of correct translations. Our hybrid SMT system processed S9f similarly to how it processed the short variant S9b: though it successfully translated the challenging element, it struggled

with the verbal contraction *couldn't* and left it in English in the target sentence. In S9g, the word *dated* unexpectedly caused problem for all of our systems, all of which interpreted the verb as meaning "to write or print the date on something" or "to find out when something old was made or formed" (first and second senses in the LDOCE entry for the verb *date*) and not "to have a romantic relationship with someone/go out with" (fourth sense in the same LDOCE entry). At first, we thought that this could have cause the mistranslation, but Google and DeepL, two systems that accurately translated *dated*, also incorrectly translated *when* as *quand* and *lorsque* respectively (i.e., the same solutions as our systems). In S9h, Google and DeepL were the only systems to pass the challenge, translating *when* as *alors que*.

Through this challenge, we found that *alors que* seems to be a safe (and general enough) translation choice for the systems, as the polysemous phrase can be used in many contexts, including for *when* meaning "in spite of the fact that". In the past, we have seen this option work for *while* expressing an opposition and *as* expressing simultaneity. We also saw that words that can be considered ambiguous to the systems (i.e., *dated*) can possibly affect how a system performs (as we have previously observed with *cooks* in S3d).

4.1.1.4. Results for with

Out of the three meanings that we tested for *with*, it was the one expressing a particular feeling or physical state (e.g., *consumed <u>with</u> guilt > rongé <u>par</u> la culpabilité*) that the systems managed to translate correctly the most. *With* expressing causality (e.g., *<u>With</u> the dog barking, I couldn't sleep > <u>Comme</u> le chien aboyait, je n'ai pas pu dormir*) came in second, while *with* meaning "in spite of" (e.g., *<u>With</u> all his debts, he still bought a car > <u>Malgré</u> toutes ses dettes, il a tout de même acheté une voiture*) did not generate any correct translations (see Table 19).

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Causality (8) (S10)	0 0.0%	0 0.0%	2 25.0%	2 25.0%	3 37.5%	2 25.0%
Feeling/state (8) (S11)	4 50.0%	2 25.0%	2 25.0%	4 50.0%	8 100.0%	8 100.0%
"*in spite of*" (8) (S12)	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Total (24)	4 16.7%	2 8.3%	4 16.7%	6 25.0%	11 45.8%	10 41.7%

Table 19 – Number and percentage of correct translations of the polysemous word *with*

Google produced the highest number of correct translations, with 11 sentences, and DeepL achieved similar results with 10 sentences. Our attention-based system was our strongest system, with six correct translations. Our hybrid SMT and RNN-based systems both produced four correct translations. Our CNN-based system was the least successful system, only producing two correct translations among all three challenges.

4.1.1.4.1. *With* expressing causality

Possible solutions for *with* expressing causality included *à cause de*, *en raison de*, *car*, *parce que*, *comme*, or use of a present participle (e.g., *I'm thinking of riding my bicycle, with the price of gas on the rise > Je pense faire du vélo, le prix de l'essence étant à la hausse*). Overall, our hybrid SMT system and our CNN-based system did not produce any correct translations, as they translated *with* as *avec* most of the time, a translation deemed too vague by Delisle and Fiola (e.g., *Je pense faire du vélo, avec le prix de l'essence à la hausse*). In fact, most of the translations that we considered incorrect were categorized as such because the systems used *avec*. The majority of the sentences (five out of the eight in total, all five of which were because of the use of *avec*) were inaccurately translated by all systems. S10b and its long variant S10e were the two sentences that the systems found easiest to translate. Interestingly enough, they were also the two sentences in which our hybrid SMT systems omitted the challenging element altogether. All in all, results for our short and long variants were very similar, with the

exception of DeepL's solution for the long variant S10g, which greatly varied from its solution in the short variant.

	"*with*" at the start of the sentence				"*with*" mid-sentence			
***With* – Causality**	**10a (S)**	**10e (L)**	**10b (S)**	**10f (L)**	**10c (S)**	**10g (L)**	**10d (S)**	**10h (L)**
Hybrid	✗	✗	✗*	✗*	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✗	✗	✗	✗
RNN	✗	✗	✓	✓	✗	✗	✗	✗
Attention	✗	✗	✓	✓	✗	✗	✗	✗
Google	✗	✗	✓	✓	✗	✓	✗	✗
DeepL	✗	✗	✓	✓	✗	✗	✗	✗

Table 20 – Results for short and long variants of sentences with with expressing causality

It was difficult for us to identify any clear and system-specific patterns that could help us explain why a system might have chosen to translate the challenging element in one way or another, because all the systems that produced inaccurate translations put forward the same, possibly "go-to" or default, solution. We were, however, able to observe certain phenomena in other parts of the sentences that highlight how the different architectures (or training data in some cases) might have played in the final translation. These will be discussed in further details in sections 4.4.1 and 4.4.2.

In our short sentences with *with* at the start of the sentence, we found error types that we also found in past challenges, though in other parts of the sentence. S10a was mistranslated by all the systems using *avec*. In S10b, our hybrid SMT system avoided the challenging element by simply omitting it from the sentence, while our CNN-based system, the only other system to incorrectly translate this sentence, used *avec*. Our RNN-based system, our attention-based system, and DeepL all accurately used *comme* as a translation. Google stood out by using a present participle to express the causality (*With everyone asked to stay home, the price of gas drastically decreased. > Tout le monde étant invité à rester à la maison, le prix de l'essence a considérablement diminué.*)

All the systems failed to accurately translate our short sentences with *with* mid-sentence, as they all used *avec* as their solution.

In our long sentences with *with* at the start of the sentence, we noted the recurring errors that we previously mentioned, such as our hybrid SMT system struggling with contractions, and our CNN-based system adding capitalization mid-sentence, but we also observe new phenomenon, which will be further discussed later on in section 4.4.2.

While all the systems failed to translate our short sentences with *with* mid-sentence, there was one correct translation in the long variants with *with* mid-sentence. In S10g, Google successfully translated *with* using a present participle. It should be noted that throughout the challenge, Google was the only system to use the present participle as a solution; all the other systems used *comme*. This shows that Google is capable of restructuring the sentence to some extent in order to express a certain meaning.

4.1.1.4.2. *With* expressing a particular feeling or physical state

In our sentences with *with* expressing a particular feeling or physical state, we anticipated solutions that generally consisted of *de* or *par*, two prepositions that work in many cases. We also accepted alternative phrases that could sound more idiomatic in certain sentences. Google and DeepL successfully translated all of our sentences in this challenge. Surprisingly, our hybrid SMT system and our attention-based system were tied in second place with the same four correct translations. Our CNN-based system and our RNN-based came in last with two correct translations each.

As mentioned previously, this was the meaning that the systems found easiest to translate out of the three meanings of *with* that we tested. We hypothesize that this could be due to the fact that these phrases tend to be quite regular, and in fact are almost formulaic or collocational (e.g., *consumed with guilt*, *burning with hatred*, *beaming with joy*, etc.). These expressions also tend to be emotion-linked words, which we are more likely to find in Google and DeepL's training data (possibly partially collected through users' usage of the tools) than our test systems' (the corpora described in Section 2.2.1). In that case, the more training data a system has, the more accurate it can be. This would explain why Google and DeepL achieved such great results in a challenge where other systems struggled to translate more than 50% of the sentences accurately. It would also explain why our hybrid SMT system came in second in terms of performance, as a statistical model would suffice for collocation-related challenges.

To express a particular feeling or physical state, *with* can only be placed mid-sentence (e.g., Her eyes were still burning *with* hatred > Ses yeux étaient toujours remplis *de* haine) and not at the beginning of the sentence (e.g., *With hatred*, her eyes were still burning.) Consequently, and similarly to what we did in S9, we only tested short and long variants, with the challenging element always found mid-sentence.

***With* – Feeling/state**	**11a (S)**	**11e (L)**	**11b (S)**	**11f (L)**	**11c (S)**	**11g (L)**	**11d (S)**	**11h (L)**
Hybrid	✗	✗	✗	✗	✓	✓	✓	✓
CNN	✗	✗*	✗	✗	✓*	✗	✓	✗
RNN	✗	✗	✗	✗	✗	✗	✓	✓
Attention	✗	✗	✗	✗	✓*	✓*	✓	✓
Google	✓	✓	✓	✓	✓	✓	✓	✓
DeepL	✓	✓	✓	✓	✓	✓	✓	✓

Table 21 – Results for short and long variants of sentences with *with* expressing a particular feeling or physical state

In our short sentences, S11a and S11b were incorrectly translated by all of our systems. All used *avec* in all the sentences, while we were anticipating more formulaic phrases (*rongé par la culpabilité*; *remplis de haine*). In S11c, although our hybrid SMT system was not the only system to successfully translate the challenging element, it was the only one to use the solution we were expecting. Our two other systems (CNN-based and attention-based) that correctly translated this sentence actually used *avec*. Normally, we would consider this to be incorrect; however, both systems removed the adjective (*beaming [with joy] > rayonnante [de joie]*), changing the phrase to *Toujours avec joie*, a phrase that would work in French (as opposed to *Toujours rayonnante avec joie*). As for our RNN-based system, it incorrectly translated the phrase as *Toujours à la joie*. In S11d, all the systems accurately translated *patients with hypothermia* as *patients atteints/souffrant d'hypothermie*. We suspect this might be due to the pair *patients with* that could be frequently found in our training data. Nevertheless, it was interesting to see that our hybrid SMT system was the only system to use *atteints de* while all the other systems used *souffrant de*.

Results for our long sentences were, for the most part, similar to the results from our short sentences. All of our systems failed to accurately translate S11e and S11f, with the majority of them using *avec* as they did in the corresponding short variants. Our CNN-based system, however, used *de* (*Il a été consommé de culpabilité*) in S11e. We marked this translation as being incorrect, as it would still require post-editing of the challenging element (We could say *consommé par la culpabilité*, though the most idiomatic structure would be *rongé par la culpabilité*). In S11g, our systems achieved slightly different results: our hybrid SMT system still managed to generate the translation that we were expecting (*radieux de joie*), while our attention-based system succeeded in translating the challenging element by going the same route as in S11c (*Toujours avec joie*), and our RNN-based system produced the same incorrect translation as it did in the short variant (*Toujours à la joie*). Our CNN-based system, on the other hand, had a different translation in the long variant. While it omitted *beaming* in the short variant, it translated it as *à faisceau* in the long (*Still beaming with joy > Toujours à faisceau avec joie*). By adding *à faisceau*, it made the translation incorrect, as post-editing would be necessary. In S11h, our CNN-based system was the only incorrect system because it translated the phrase *patients with* literally, as *patients avec*. Most of the systems translated *patients with* as *patients souffrant de*, with the exception of our hybrid SMT system, which once again had a different solution than the rest. Its solution in the long variant (*présentant des*) was even different than its solution in the short variant (*atteints de*). We hypothesize that our hybrid SMT system, as a PBSMT system, is likely to draw heavily on immediate context, and since the words following *with* in the long variant (*patients with thyroid problems or severe hypothermia*) were different than the words following *with* in the short one (*patients with hypothermia*), it produced different translations.

4.1.1.4.3. *With* meaning in spite of

With meaning "in spite of" was incorrectly translated by all systems, in all sentences. This could be because this meaning requires the systems to establish a logical link between the content of the two phrases linked by *with*. We believe that the link is slightly more abstract and subjective than *with* expressing causality. Furthermore, *with* meaning

"in spite of" is a meaning found less frequently than the other two meanings we tested. In the LDOCE, *with* expressing causality appears as the fourteenth sense in the entry for *with* and *with* expressing a particular feeling or physical state as the fourth one, while *with* meaning "in spite of" is the nineteenth sense.

***With* – "in spite of"**	**12a (S)**	**12e (L)**	**12b (S)**	**12f (L)**	**12c (S)**	**12g (L)**	**12d (S)**	**12h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✗	✗	✗	✗
RNN	✗	✗	✗	✗	✗	✗	✗	✗
Attention	✗	✗	✗	✗	✗	✗	✗	✗
Google	✗	✗	✗	✗	✗	✗	✗	✗
DeepL	✗	✗	✗	✗	✗	✗	✗	✗

Table 22 – Results for short and long variants of sentences with *with* meaning "in spite of"

With meaning "in spite of" was consistently translated as *avec* in all of our sentences, regardless of their length or of the location of *with* within the sentences.

4.1.2. Asymmetrical equivalence – Homographs

In addition to ambiguities in the source language, we also tested our systems for lexical difficulties stemming from asymmetrical equivalence. In this case, we were specifically looking at homographs, i.e., words that share the same spelling but do not share the same meaning. As seen in Table 23, DeepL performed better than most for homographs.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Desk/*Office* (2) S13a & S13e	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Appreciate/*Enjoy* (2) S13b & S13f	2 100%	0 0.0%	1 50.0%	0 0.0%	0 0.0%	1 50.0%
Like/*Love* (2) S13c & S13g	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 50.0%
Cool/*Fresh* (2) S13d & S13h	0 0.0%	0 0.0%	0 0.0%	1 50.0%	0 0.0%	2 100.0%
Total (8)	2 25.0%	0 0.0%	1 12.5%	1 12.5%	0 0.0%	4 50.0%

Table 23 – Number and percentage of correct translations of homographs

For this challenge, we anticipated translations where the systems would be able to avoid a repetition in the target, either by using a synonym (e.g., The company bought new *desks* for all of their *offices* > La compagnie a acheté de nouveaux *bureaux* pour tous ses *immeubles*) or by finding a different turn of phrase (e.g., The wine was delicious, *cool and fresh* and fruity > Le vin était délicieux: *frais, d'un goût rafraîchissant* et fruité). A wrong translation would include a repetition that would sometimes make the sentence awkward (e.g., Le vin était délicieux: *frais et frais* et fruité), or illogical (While I *appreciate* the comments, I don't particularly *enjoy* them > Bien que j'*apprécie* les commentaires, je ne les *apprécie* pas particulièrement). Surprisingly, this is one of the rare challenges where some of our systems managed to render correct translations while Google failed to correctly translate any sentence. Our CNN-based system was also unable to produce any accurate translation. Although our hybrid SMT system came in first among our systems in terms of performance, it only generated two correct translations, while our RNN-based system and attention-based system both had one correct translation.

We only tested for short and long variants in this challenge, as other than their relative proximity to one another, the position of the homographs in the sentence does not really matter, considering that they do not play a structural role. It should also be noted that some of these sentences are not incorrect translations in and of themselves, they are simply sentences that would require post-editing by a human translator because they are

written rather awkwardly (e.g., *L'entreprise a acheté de nouveaux bureaux pour tous ses bureaux.*).

Homographs	13a (S)	13e (L)	13b (S)	13f (L)	13c (S)	13g (L)	13d (S)	13h (L)
Hybrid	✗	✗	✓	✓	✗	✗	✗*	✗*
CNN	✗	✗	✗	✗	✗	✗	✗	✗
RNN	✗	✗	✓	✗	✗	✗*	✗	✗
Attention	✗	✗	✗	✗	✗	✗*	✓*	✗
Google	✗	✗	✗	✗	✗	✗*	✗	✗
DeepL	✗	✗	✗	✓	✓	✗*	✓*	✓

Table 24 – Results for short and long sentences containing homographs

There were more correct translations in our short sentences than in our long variants. In our short sentence, S13a was mistranslated by all systems. All the systems failed to translate S13a and its long variant, S13e, accurately. S13g was also problematic for all systems.

In S13a, all the systems translated *desks* and *offices* as *bureaux* and *bureaux*. A human translator would have resorted to another word for *bureaux*, such as *immeubles* or *édifices*, to avoid repetition or would have to postedit to get a usable output. In S13b, we tried to contrast the two feelings expressed by *appreciate* and *enjoy*, to emphasize how correct or incorrect a translation was, but also to provide the systems with a "fair" opportunity to differentiate the two words. To do so, we used *while* expressing an opposition, in addition to contrasting the feelings expressed in the two propositions with negation in the second proposition (*While I appreciate the comments people leave on my videos, I don't particularly enjoy them*). Our hybrid SMT system correctly translated the two words as *apprécie* and *profiter*, and our RNN-based system also correctly used *apprécie* and *aime*. On the other hand, our CNN-based system and our attention-based system, as well as Google and DeepL, awkwardly used *apprécie* in both propositions (*Bien que j'apprécie* […], *je n'apprécie pas*…). In S13c, DeepL was the only system to accurately translate the challenging element. Once again, we tried to emphasize the difference between our homographs' meaning in our source sentence to highlight any mistranslation. In our sentence, our subject hesitates between two feelings, *like* or *love*.

We found that most systems struggled with translating *like* and interpreted the word as being a conjunction (i.e., meaning as though) rather than a noun derived from the verb. We do recognize, however, that this is quite an unusual use, which can cause problems for any system. Indeed, our hybrid SMT system and our CNN-based system translated *like or love* as *comme ou l'amour*, our attention-based system, as *comme ou si il aimait*, and Google, as *comme ou si c'était de l'amour*. As for our RNN-based system, it generated the incorrect translation that we were expecting of systems should they fail to translate the challenge (i.e., it repeated the same word in the target sentence: *aimé ou aimé*). Contrastingly, DeepL managed to move away from the original structure found in the source and generate *de l'amour ou non*. We hypothesize that this could be because DeepL is based on the Linguee database, which consists of human translations. In these human translations, sentences containing homographs are most likely reworded so when fed into DeepL's the training data, similar cases do not contain repetitive structures, allowing the systems produce more fluid/less repetitive sentences. In S13d, the two systems (attention-based and DeepL) that passed the challenge had the same solution (*cool and fresh > frais*). Our hybrid SMT system failed the challenge because it kept[41] one of the two words in the source language (*cool and fresh > cool et frais*), while our CNN-based and attention-based systems, as well as Google, failed because they repeated the word (*frais et frais*), again making post-editing necessary.

Results for our long sentences were not quite as good as for our short variants. In some sentences, such as S13e, the translations produced by the systems were similar to the solutions they produced in the corresponding short variant. In others, we found new solutions that were sometimes correct alternatives, sometimes wrong translations. For example, in S13f, our hybrid SMT system produced a correct translation, but different than the one it came up with in the short variant S13b. In S13b, it translated *appreciate* and *enjoy* as *apprécie* and *profiter*, whereas in S13f, it translated it as *apprécie* and *aiment*. Still in S13f, our RNN-based system, that had previously accurately translated S13b, failed to translate the challenge, repeating *apprécie* in a phrase that would certainly require human post-editing (*j'apprécie mais je n'apprécie pas*). Conversely, DeepL, one

[41] We marked our hybrid SMT system's translation of S13d as being difficult for us to assess because we do not know if *cool* has been left in the source language or if it has been translated as an anglicism.

of the systems that failed to translate S13b, found a solution that works in S13f. In S13f, DeepL used *j'apprécie, mais pas particulièrement*, thus avoiding any awkward repetition. S13g was another sentence that all systems failed to translate accurately. We were able to identify how our hybrid SMT system and our CNN-based system mistranslated *like* and *love*, but we were not able to find both elements in our other translations. Our RNN-based system only translated *like* as *comme* and omitted to translate *love*. Google and DeepL, on the other hand, translated *love* as *aimait*, but omitted *like*. Our attention-based system did not translate either of the two elements. In S13h, our hybrid SMT system produced an incorrect translation comparable to the one it produced in S13d (containing the English form *cool*). Our attention-based system failed the challenge, as it repeated *frais* in the long variant, something it did not do in the short version. DeepL was successful in avoiding repetition, but it found a different solution than it did in the short variant. In S13d, it translated *cool and fresh* as the encompassing *frais*, but in S13h, it translated it as *frais, puissant*. Although *puissant* might not be the accurate translation for *fresh*, this translation was deemed correct, since the purpose of this challenge was, first and foremost, to see if a system could avoid repetition. Furthermore, *puissant* is an adjective that could be used to describe wine: If we look at the target sentence alone, the part that corresponds to the challenging element would not require post-editing.

All in all, we noticed through this challenge that homographs are hard for MT systems to translate, not because they are unable to find a translation for the source, but because they might be unable to detect what we humans consider awkward wordings. In sentences where repetitions disrupt logic (i.e., S13b, S13f), errors made by the systems could be attributed to their inability to build a logical context from inference. In sentences like these, instead of building a context around the meaning of the phrases, they would have to build a context around the meaning of the clauses, not only identifying the right linkages to make between them, but also ensuring that the words within one clause do not contradict those in the other. We also found that ambiguous words (such as *like* used in the context of S13c and S13g) can lead to translations that are "more erroneous" (i.e., that require more post-editing).

4.2. Results – Syntactic difficulties

Syntactic difficulties rendered results that were slightly more homogeneous. Surprisingly, our hybrid SMT system performed better than our three NMT systems, with 45% of the sentences correctly translated. Our RNN-based system came in second, while our CNN-based system and our attention-based system tied in last place with only 20% of their sentences accurately translated. Google and DeepL scored higher.

Table 25 below gives an overview of the results obtained from our experimental and reference systems for our syntactic challenges.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Scope (S14-S15)	7 43.8%	2 12.5%	3 18.8%	3 18.8%	10 62.5%	10 62.5%
Anaphora (S16-S18)	11 45.8%	6 25.0%	12 50.0%	5 20.8%	14 58.3%	16 66.7%
Total	18 45.0%	8 20.0%	15 37.5%	8 20.0%	24 60.0%	26 65.0%

Table 25 – Number and percentage of correct translations for syntactic difficulties

Our hybrid SMT system struggled with the anaphora with the pronoun *they* the most, whereas our NMT systems found the scope of modifiers the hardest to translate. Indeed, our CNN-based system did not translate any of the scope-of-modifier sentences correctly, and our RNN-based and attention-based systems only accurately translated one of the sentences.

4.2.1. Ambiguity – Scope

When we tested the systems for scope-related difficulties, we looked at scope of modifiers and scope of conjunctions. In both cases, we checked whether the systems could make the right agreements in long and short sentences.

As shown in Table 26, results for our scope of conjunction examples were, for the most part, stronger than the results we achieved in our scope of modifier examples.

Google was the only system to score lower for scope of conjunctions than for scope of modifiers.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
Scope of modifier (S14)	3 37.5%	0 0.0%	1 12.5%	1 12.5%	7 87.5%	5 62.5%
Scope of conjunction (S15)	4 50.0%	2 25.0%	2 25.0%	2 25.0%	3 37.5%	5 62.5%
Total	7 43.8%	2 12.5%	3 18.8%	3 18.8%	10 62.5%	10 62.5%

Table 26 – Number and percentage of correct translations per category of scope

Overall, Google and DeepL led in terms of performance, with both successfully translating 62.5% of our scope-related sentences (although despite this overall comparability, their performance varied quite a bit in the different types of scope challenges tested, as discussed below). Our hybrid SMT system came in second, followed by our RNN-based and attention-based systems, and our CNN-based system came in last.

We hypothesize that our hybrid SMT system got higher scores than our neural systems because association of a modifier with the correct noun can be achieved through a statistical model. Statistical models partly rely on the number of occurrences of a translation in their training data to reach the most likely translation candidate. As such, their context window might be limited because they are sometimes simply juxtaposing chunks (i.e., occurrences of a certain phrase) without truly analyzing the relationship between the chunks or making linkages. While this is often seen as a weakness, in some cases, such as when dealing with scopes, it can benefit a system. For example, in the example *The only Australian wine lovers I know are coincidentally also Australian*, while an NMT system might analyze the phrase *Australian wine lovers* as one block, a statistical model might, instead, view the phrase *Australian wine* and the word *lovers* as two separate sections, if that is how the sections are found in the training data. In other words, if parts of our challenging element have appeared before in the training data, it is easier for our hybrid SMT model to retrieve it as a phrase than for our neural systems to build a context around it to determine which noun would be the most likely translation.

4.2.1.1. Scope of modifier

In our sentences with ambiguity stemming from scope of modifiers, we created sentences where a modifier would be followed by two nouns. A translation is deemed correct if the system associates the modifier with the right noun. While a human would identify the cues found elsewhere in the sentence or use logic (in cases where the modifier can logically only be associated to one of the two nouns), a system would instead, ideally, determine the solution from the word embeddings and associations found in the broader context, or that would normally not be found if they are improbable.

Results for our neural systems were bleak, with our CNN-based system failing to correctly translate any of our sentences and our RNN-based and attention-based systems only successfully translating one. Our hybrid SMT system had the highest number of correct translations, with only three. There was a noticeable difference in terms of performance between our trained-in-house systems and Google and DeepL, which performed much better.

Scope of modifiers	14a (S)	14e (L)	14b (S)	14f (L)	14c (S)	14g (L)	14d (S)	14h (L)
Hybrid	✓	✓	✗	✗	✗	✓	✗	✗
CNN	✗	✗	✗	✗	✗	✗	✗*	✗
RNN	✗	✗	✗	✗*	✗	✗	✓	✗*
Attention	✗	✗	✓	✗*	✗	✗	✗	✗*
Google	✓*	✓*	✓	✓	✓	✗	✓	✓
DeepL	✗	✓	✗	✗	✓	✓	✓	✓

Table 27 – Results for short and long variants of sentences tested for scope of modifiers

The two sentences that our neural systems were able to accurately translate were short. Our hybrid SMT system only correctly translated one short sentence, while Google correctly translated all four, and DeepL, two. In S14a, our hybrid SMT system was the only one to produce the translation that we were expecting (i.e., *amateurs de vin australien*). The other system that got this sentence right, Google, did so by using the plural everywhere (*amateurs de vins australiens*); the plural makes it impossible to tell to

which noun the modifier had been linked by the system (or, in fact, if it had been linked to either one). We marked this translation as correct, though we cannot draw any conclusion from it or formulate any hypothesis based on it. In S14b, our attention-based system and Google correctly translated the challenging element (*dehydrated dog treat*) by associating *dehydrated* with *treat* and not *dog*. Our CNN-based omitted the word *treat* altogether and translated *you lost the dehydrated dog treat* as *vous avez perdu le chien déshydraté*. We also note that, in this sentence, all of our systems that translated the word *treat* interpreted it as meaning *traitement* or *traiter*, rather than *gâterie*. This could have influenced their resulting translation, but our attention-based system still managed to associate *dehydrated* with the right noun (i.e., not *dog*), despite using *traitement* (*vous avez perdu le traitement déshydraté des chiens*). In S14c, only the commercial systems generated correct translations. Our systems all associated *new* with *window* in *new arrivals window display* and translated the phrase as *nouvelle fenêtre*. This sentence in particular was likely harder for systems to translate because the modifier was followed by not two, but three nouns. This is something we noted and will take into consideration in the future,[42] and S14c is among the sentences that we would revise should the opportunity arise. In S14d, a weakness from our CNN-based system that we have identified in the past caused the system to mistranslate the challenging element: the system split the sentence, but did so mid-challenge. As a result, it translated *my abnormally crisp asparagus story* as *mes anormalement criants: L'histoire de l'épereur*. Once again, we find the colon used to introduce an example or explanation, which is widely used in French, but is superfluous and inaccurate in this case. In addition, we also found that our neural systems struggled with the word *crisp*, which, considering our limited training data, could have been a rare word. (Nevertheless, our hybrid SMT system did not have any issue translating this item, indicating that, while it is probably not common, it is not absent either.)

Our NMT systems failed to translate any of our long variants accurately. Our hybrid SMT system produced two correct translations, one of which is S14e, corresponding to S14a, the only short sentence our hybrid SMT system had accurately

[42] For more discussion of possible changes to the challenge set for future use, see Section **Error! R eference source not found.** or refer to our Conclusion

translated. Google and DeepL were also among the systems that correctly translated S14e. Google used the solution it used in S14a, avoiding all risks of errors by pluralizing all the elements (*amateurs de vins australiens*), whereas DeepL used the same solution as our hybrid SMT system, which was the solution that we were expecting (*amateurs de vin australien*). It was interesting to see DeepL fail our challenge in the short variant, but succeed in the long one, as we would have expected the longer sentence to be harder for the system to translate. Results for S14f were very different from the results we got in the corresponding short variant. In S14f, Google was the only system to successfully associate *dehydrated* with *treat* and not *dog* (*dehydrated dog treat > gâterie déshydratée pour chien*). Our CNN-based system and DeepL associated *dehydrated* with *dog*, while our attention-based system produced a translation that made it impossible for us to determine whether the system associated *dehydrated* with *dog* or *treat* (*traitement de chien déshydraté*). As for our hybrid SMT system and our RNN-based system, they also produced translations where it was impossible for us to determine whether the systems had made the right association or not. Our hybrid SMT system made an incorrect agreement for the only noun in the immediate context (*chien déshydratés traiter*) and our RNN-based system produced a translation so far from the source that we could not associate the elements of the target with those of the source (*dehydrated dog treat > déshydratable et déshydraté*). In S14g, it was interesting to see that our hybrid SMT system, which had previously mistranslated the short version of the sentence (*new arrivals window display > nouvelle fenêtre d'affichage des arrivées*), correctly translated the long variant (*vitrine de nouveaux arrivants*). In contrast, Google, which succeeded in the short version, failed the long version (*nouvelle vitrine des arrivées*). It is also worth noting that although DeepL successfully translated both the short and long variants, it came up with a different solution in each case (*vitrine pour les nouveaux arrivant*s in the short and *vitrine des nouveaux arrivages* in the long). In S14h (*abnormally crisp asparagus story*), the commercial systems were the only systems to accurately translate the challenge. Our hybrid SMT system and our CNN-based system incorrectly associated *crisp* with *story* rather than *asparagus*. Our RNN-based and attention-based systems made errors in the agreements, which once again made it impossible for us to determine

to which noun the systems meant to associate the modifier (RNN: *histoire d'asperges anormalement criant*; Attention: *histoire d'asperges anormalement cristé*).

4.2.1.2. Scope of conjunction

In sentences where we tested for scope of conjunctions, we had short and long sentences containing noun phrases made up of two nouns linked by a conjunction, accompanied by a modifier. As explained in 3.4.1.2, the modifier we selected can logically only be associated with one of the two nouns (e.g., *solar-powered calculators and pencils*, where the only logically possible grouping is *[solar-powered calculators] and pencils*). We used nouns linked by coordinating conjunctions, such as *and* and *or*, and created sentences where the modifier can logically only be associated with the first noun. This allowed us to determine whether the systems were able to interpret the scope of the conjunctions as intended, or if they applied the modifer either to both nouns (in which case it would appear in plural form) or to the wrong one (in which case the modifier would be singular, but placed after the second noun).

Scope of conjunction	15a (S)	15e (L)	15b (S)	15f (L)	15c (S)	15g (L)	15d (S)	15h (L)
Hybrid	✓	✓	✗*	✓	✗	✓	✗*	✗*
CNN	✗	✗	✓	✓	✗*	✗*	✗*	✗*
RNN	✗	✗	✗*	✗*	✗*	✗*	✓	✓
Attention	✗	✗	✓	✓	✗*	✗*	✗*	✗
Google	✓	✓	✗	✗	✗	✓	✗	✗
DeepL	✓	✓	✓	✗	✓	✓	✗	✗

Table 28 – Results for short and long variants of sentences tested for scope of conjunction

We saw an overall even performance across the sentences, as five of the eight sentences (S15a, S15b, S15e, S15f, and S15g) were correctly translated by three systems—although they were different systems depending on the sentence. Three sentences (S15c, and S15d and its long variant S15h) were correctly translated by only one system. Interestingly, our NMT systems all accurately translated two sentences each, sentences that corresponded to a short and long variant pair. Our hybrid SMT system

showed the strongest performance out of our experimental systems, with 50% of the sentences translated accurately, the majority of which were long sentences. DeepL had the largest number of sentences correctly translated, while Google had one fewer than our hybrid SMT system.

Most systems only correctly translated one of our short sentences, with the exception of DeepL, which had three correct translations. Surprisingly, a number of the mistranslations were due to systems completely omitting to translate the modifier, rather than incorrectly representing the scope of the conjunction. In S15a, all of our NMT systems mistranslated the sentence by adding the modifier after the wrong noun. Our hybrid SMT system, as well as Google and DeepL, made the correct agreement, which indicates the logical association of the modifier (*solar-powered*) with the first noun (*calculators*) and not the second (*pencils*). They all translated the phrase by inserting the modifier mid-phrase, after the first noun only (*solar powered calculators and pencils > calculatrices solaires et des crayons*). In S15b, our CNN-based system, our attention-based system, and DeepL accurately translated the phrase *itchy skin or wood*. Although our CNN-based system mistranslated the phrase *itchy skin* as *peau de démangeaison* (instead of *peau qui démange*), it did identify the correct noun (*peau*) and associated the modifier with it. Our attention-based system also struggled with translating *itchy* and left it in the target sentence (*la peau d'itchy ou le bois*), but like our CNN-based system, it identified the correct noun. Our hybrid SMT system translated *itchy skin* as *démangeaisons de la peau*, which, in and of itself, is not an entirely wrong translation. The reason why this sentence was marked as being incorrect is because it was impossible for us to determine[43] whether the system had associated the modifier with just one noun, or both, as it used an indefinite article for *démangeaisons* (*des démangeaisons de la peau*). As such, the phrase can be interpreted as *des démangeaisons de la peau ou des démangeaisons du bois*, or as *des démangeaisons de la peau, et du bois*. Our RNN-based system also failed the challenge, but by omitting to include the modifier in the target sentence. Google was the only system to produce the wrong translation that we were expecting, by grouping the two nouns (*la peau ou le bois qui démange*). In S15c, DeepL

[43] For more information about our evaluation method, see section 3.2

was the only system to successfully translate the challenging element, despite the mistranslation of the modifier itself (*fluffy dogs and geckos > des chiens en peluche et des geckos*). In this context, we consider the translation as correct, as we were focusing on the issue of scope. Our hybrid SMT system and Google grouped the two nouns, while all of our NMT systems left the modifier out of the target sentence. It should also be noted that our hybrid SMT system failed to make the right gender agreement, as *chiens* and *geckos* are both masculine nouns and it used *duveteuses* (*les chiens et les geckos duveteuses*). In S15d, we saw a similar pattern, where, aside from our RNN-based system, all of our systems failed the challenge because they all omitted the modifier from the translation. Our RNN-based system was the only successful system, outperforming even Google and DeepL with its translation (*wood blinds and curtains > des stores en bois et des rideaux*). Google and DeepL both made the same mistake of grouping the two nouns and associating the modifier to the grouping (*des stores et des rideaux en bois*).

As mentioned previously, results in our long sentences were very similar for our NMT systems, but differed for our hybrid SMT, as well as for Google and DeepL. Results for S15e were identical to the ones for the corresponding short variant S15a. Our hybrid SMT system, Google, and DeepL were successful in identifying that *solar-powered* should only be associated with *calculators* (*solar-powered calculators and pencils*), whereas our NMT systems misinterpreted the phrase as *calculatrices et crayons à énergie solaire*. In S15f, we marked our hybrid SMT system's translation as correct because of the articles the systems used: in the short variant, it used an indefinite article for both nouns, while in this long variant, it used *des démangeaisons de la peau ou le bois*. This makes the interpretation unequivocal as French requires the repetition of prepositions placed before a series of nouns (i.e., had *démangeaisons* been associated with both nouns, the system would have produced *des démangeaisons de la peau et du* [i.e., *de le*] *bois*). By using a definite article for *bois*, it separated the noun from *des démangeaisons* and the translation can be interpreted as *pour hydrater des démangeaisons de la peau et pour hydrater le bois*. For the most part, the other systems produced similar translations in S15f as they did in the corresponding short variant S15b, with the exception of DeepL. In S15b, it successfully translated the challenge, but in S15f, it produced the same inaccurate translation as Google (*la peau ou le bois qui*

démange). In S15g, our hybrid SMT system and Google produced correct translations, while they had failed to do so in the short variant. The modifier's location changed within the phrase in both cases. Our hybrid SMT went from *les chiens et les geckos duveteuses* in S15c to *les chiens duveteuses et les geckos* in S15g; Google went from *des chiens et des geckos moelleux* in S15c to *des chiens moelleux et des geckos*. In S15h, all the systems produced similar results as they did S15d and our RNN-based system was the only system to correctly identify that *wood* could only be associated with *blinds* and not *curtains* (*my wood blinds and curtains > mes stores de bois et mes rideaux*).

4.2.2. Ambiguity – Anaphora

Anaphora are known problems for MT systems (Guillou & Hardmeier, 2016; Hardmeier & Guillou, 2018), as they require the correct processing and linking with the antecedent in context in order to be accurate. We chose three pronouns to test: *it*, *they*, and *these*. As seen in Table 29, *they* was the pronoun the systems struggled to translate the most, with only 14 correct translations across all six systems, while *it* and *these* each generated 25 correct translations.

	Hybrid SMT	CNN	RNN	Attention	Google	DeepL
It (S16)	6 75.0%	3 37.5%	5 62.5%	2 25.0%	3 37.5%	6 75.0%
They (S17)	0 0.0%	1 12.5%	2 25.0%	3 37.5%	4 50.0%	4 50.0%
These (S18)	5 62.5%	2 25.0%	5 62.5%	0 0.0%	7 87.5%	6 75.0%
Total	11 45.8%	6 25.0%	12 50.0%	5 20.8%	14 58.3%	16 66.7%

Table 29 – Number and percentage of correct translations per category of anaphora

Overall, DeepL achieved the highest score in terms of performance, followed by Google. Our RNN-based system followed close behind and our hybrid SMT system came in fourth. Our CNN-based system achieved much lower results, and our attention-based system unexpectedly came in last.

4.2.2.1. Results for it

In sentences where we tested for *it*, we looked at agreement in gender, in short and long variants, with interruption (i.e., another noun occurring between the pronoun and antecedent) (e.g., *The recital was tonight but there were issues with the stage so it was cancelled*) and without (e.g., *I wanted the bag with the side pocket because it adds a lot of space*).

	Without interruption				With interruption			
Anaphora – *it*	**16a (S)**	**16e (L)**	**16b (S)**	**16f (L)**	**16c (S)**	**16g (L)**	**16d (S)**	**16h (L)**
Hybrid	✓	✓	✓	✓	✗	✗	✓	✓
CNN	✓*	✗	✗	✗	✗	✗	✓	✓
RNN	✓	✓*	✗	✗*	✗	✓	✓	✓
Attention	✗*	✗*	✗	✗	✗	✗	✓	✓
Google	✗	✓	✗	✗	✗	✗	✓	✓
DeepL	✓	✓	✗	✗	✓	✓	✓	✓

Table 30 – Results for short and long variants of sentences with the anaphora *it*

Results for long and short variants were similar, but it was interesting to see that there were slightly more correct translations in the sentences with interruption than without. We were expecting interruptions to add a layer of difficulty to the challenge; however, 62.5% of the sentences with interruption were correctly translated, but only 41.7% of those without interruption. Overall, the sentence pair (short and long) that the systems struggled to translate the most was S16b and S16f, and the sentence pair that produced the highest number of correct translations was S16d and S16h.

Our hybrid SMT system was the only system to correctly translate both of our short sentences without interruption. In S16a, our hybrid SMT system inaccurately translated the noun to which *it* refers (it translated *mint*, the plant, as *monnaie*), but accurately translated it as the feminine pronoun *elle* that we were looking for, as both *menthe* and *monnaie* are feminine. Our CNN-based and RNN-based systems also accurately translated *it* as *elle*, however both systems translated *my mint* as *mon menthe*. This unexpected discrepancy and made it impossible for us to determine whether the systems had truly interpreted *menthe* as being feminine or not, since it used a masculine

article with the (normally feminine) noun, but accurately used a feminine pronoun to refer to it. Our attention-based system surprisingly translated *my mint* as *mon minet* and, as a possible consequence, translated *it* as a masculine pronoun, perhaps to reflect the first half of the sentence. Google had the incorrect translation that we were expecting: the system accurately translated the first half of the sentence, but translated *it* as *il*, likely because it associated *it* with *un pot* rather than *ma menthe*. DeepL was the only system to accurately translate both the pronoun and the noun to which *it* referred. In S16b, our hybrid SMT system was the only system to correctly translate *it* as *elle*, referring to the side pocket rather than the bag (*I wanted the bag with the side pocket because it adds a lot of space. > Je voulais le sac avec la poche de côté, car elle ajoute beaucoup d'espace.*). All the other systems, though they accurately translated the noun as *la poche*, mistranslated *it* as *il*.

In our short sentences with interruption, S16c was mistranslated by most systems, while S16d was correctly translated by all. In S16c (e.g., He put his brother's *photo* in a book after breaking the frame *it* was in), almost all the systems mistranslated *it* as the masculine *il*, while the pronoun was referring to the feminine noun *photo*. We believe this could be in part due to the fact that we accidentally inserted two interruptions (*book* and *frame*) instead of just one as we had originally intended. This most likely made it harder for the systems to link *it* to the appropriate noun. DeepL was the only system to pass this challenge by translating *it* as *elle* (*He put his brother's photo in a book after breaking the frame it was in. > Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequel elle se trouvait*.). In S16d, all the systems accurately translated the challenging element *it* as *il*, representing the masculine noun *récital*. It is worth mentioning that, though there were more interruptions in S16c, the distance between the noun and the pronoun in S16d was greater in terms of word count (there are ten words separating the noun and the pronoun in S16d, as opposed to seven words in S16c). This could suggest that ambiguity is a greater challenge to systems than sentence length.

In our long sentences without interruption, our hybrid SMT system, our RNN-based system, our attention-based system, and DeepL performed similarly in S16e and in the short variant S16a. Our CNN-based system, however, translated *my mint* as *mon cuir*, then later translated *it* as the feminine *ell*e, although *mon cuir* is a masculine noun. We

marked this translation, with its faulty agreement, as incorrect.[44] Google also performed differently in S16e than it did in the short variant S16a, in that it accurately translated *it* as *elle* in the long variant. Results for S16f were similar to the results we got in the short variant S16b. The only noticeable difference was that our RNN-based system mistranslated *I wanted the brown leather bag* as *Je voulais que le sac de cuir brun* (literally *I wanted that the brown leather bag*). Since the system changed the overall structure of the sentence, it omitted the challenging element in the target (*Je voulais que le sac de cuir brun avec la pochette latérale blanche ajoute beaucoup d'espace*, literally *I wanted the brown leather bag with the white side pocket to add a lot of space*), and was therefore considered incorrect.

In our long sentences with interruption, results were similar to those for the short variants, with the only difference being our RNN-based system's translation of S16g. In the short variant, the system had failed the challenge, but it successfully translated *it* as *elle* in the long one (*He put his little brother's graduation photo in a book with all of their childhood photos after accidentally breaking the wooden frame it was in. > Il a mis la photo de son petit frère dans un livre avec toutes leurs photos d'enfance après avoir rompu accidentellement le cadre de bois qu'elle était*).

Through this first challenge involving anaphora, we found that sentence length might not influence the systems' performance as much as we had hypothesized. Instead, elements causing ambiguities in the sentence appear more likely to affect a system's translation. This possibility will be explored further in section 4.3.1.

4.2.2.2. Results for they

Of our three challenges involving anaphora, our hybrid SMT system found *they* the most problematic to translate. This is interesting to us, as we originally believed that *these* would be the most difficult to translate, based on the additional element (proximity) the systems had to take into consideration.

[44] In previous examples where the noun was mistranslated, a feminine noun had still been used, making the agreement consistent with the rest of the sentence. In those cases, we were able to mark the sentences as correct, since we were focusing on the correct gender agreement.

	Cue before pronoun				Cue after pronoun			
Anaphora - *they*	**17a (S)**	**17e (L)**	**17b (S)**	**17f (L)**	**17c (S)**	**17g (L)**	**17d (S)**	**17h (L)**
Hybrid	✗	✗	✗	✗	✗	✗	✗	✗
CNN	✗	✗	✗	✗	✓*	✗	✗	✗
RNN	✗	✗	✓	✗*	✗	✗	✓	✗*
Attention	✗	✓*	✗	✗*	✗	✗	✓	✓
Google	✓*	✓	✓	✓	✗*	✗*	✗	✗
DeepL	✓*	✓*	✓	✓	✗*	✗*	✗	✗

Table 31 – Results for short and long variants of sentences with the anaphora *they*

Overall, the results were slightly better in our short sentences (33.3%) than in our long sentences (25%), though they were not high in both cases. Google and DeepL accurately translated the same two sentences in the short variants and also in their corresponding long variants, accurately translating 50% of the sentences overall. Our attention-based system came in second, followed by our RNN-based system, and our CNN-based. Our hybrid SMT system did not manage to accurately translate any sentence.

For sentences with *they*, we were looking for translations where the systems identified not only the right gender agreement, but also the right number agreement. We developed sentences where the cue indicating whether the pronoun should be feminine or masculine was placed either before or after the pronoun. Sentences where the cue was placed before the pronoun produced more correct translations than the ones with the cue placed after the pronoun, although our sample might be too small for us to draw any solid conclusions (ten correct translations with cue placed before the pronoun vs. four correct translations with cue placed after the pronoun).

In our short sentences with the cue before the pronoun, Google and DeepL were the only systems to accurately translate both sentences. In S17a, they passed the challenge by avoiding the use of any pronoun, which worked in this case (*The matriachs are the rulers of the family and they usually make important decisions. > Les matriarches sont les [dirigeants/chefs] de la famille et prennent généralement [des/les] décisions importantes.*). All the other systems failed because they translated *they* representing the *matriarchs* as *ils*. We also noticed that all the systems translated *rulers* using a masculine

noun (hybrid SMT: *gouvernants*; CNN-based, RNN-based, attention-based and Google: *dirigeants*, DeepL: *chefs*), while *rulers* stood for *matriarchs*. In S17b, in addition to Google and DeepL, our RNN-based system also generated an accurate translation (*The waitresses came with the bill, but they forgot to include some of our drinks. > Les serveuses sont arrivées avec la facture, mais elles ont oublié d'inclure certaines de nos boissons.*). Our attention-based system was the only system to translate *waitresses* as *serveurs*, rendering a grammatically correct sentence, but an inaccurate translation.

While Google and DeepL achieved good results in our short sentences with the cue placed before the pronoun, they struggled to translate our short sentences where the cue was placed after the pronoun. In S17c, our CNN-based system was the only system to accurately translate the challenge, but it did so by translating the strictly masculine noun *nurses* in this context as *infirmières* et *infirmiers*. Consequently, the system translated *they* as *ils*. While the translation is deemed correct in the context of this challenge, it did not allow us to interpret the inner workings of the system since it still included *infirmières* in a sentence where we discuss *paternity leave*. Google and DeepL used the same solution in S17c as they did in S17a, only this time, the translation without the pronoun did not work because they used the feminine noun *infirmières* at the beginning of the sentence (*The nurses are unhappy they got an extra shift after coming back from paternity leave. > Les infirmières sont mécontentes d'avoir [eu un quart de travail/obtenu un poste] supplémentaire après leur retour de congé de paternité.*). The other systems (hybrid SMT, RNN-based, and attention-based) failed the challenge by translating *nurses* as *infirmières* and thus, *they* as *elles*, not taking into consideration the cue *paternity leave* placed after the pronoun. It is worth noting, however, that the two cues after the pronoun are also in a word other than the antecedent (*paternity leave*, *all-male team*), instead of in the antecedent itself (as was the case for *matriarchs*, *waitresses*), which could have affected the ease of the analysis. To determine the right gender agreement, systems will typically find the pronoun's antecedent to generate the most likely pronoun. In our sentences with the cue before the pronoun, as stated above, the cue was the antecedent itself, making the analysis process more direct than in our sentences with the cue after the pronoun, where the antecedents we chose were gender neutral, forcing the systems to look for cues elsewhere in the sentence. In S17d, our RNN-based system and our

attention-based system both successfully translated the sentence by using the feminine plural noun *personnes* to designate *those two*. We believe this is a happy coincidence that allowed the systems to translate *they* as *elles*, even though they may not have interpreted *those two* as being females using the context provided (*When those two joined that team, they turned the all-male team into a mixed one. > Lorsque ces deux personnes se sont jointes à [l'équipe/cette équipe], elles ont transformé l'équipe [de tous les hommes/entièrement masculine] en une équipe mixte.*). Our hybrid SMT system translated *those two* as *ces deux* and was not able to accurately translate *they* as *elles*. Google and DeepL came up with a similar solution, where they translated *those two* as *ces deux-là* and mistranslated *they* as *ils*. As for our CNN-based system, it mistranslated *those two joined that team* as *ces deux équipes se sont jointes à cette équipe*. What is interesting is that, with this mistranslation, it should have translated *they* as *elles* because *équipes* is feminine plural, but it still failed to render the correct agreement and translated *they* as *ils*.

In our long sentences with the cue before the pronoun, Google and DeepL once again successfully translated both sentences, however there were small variations in our other systems' results. In S17e, our attention-based system that originally failed to translate the short version of the sentences, correctly translated the long variant using the solution that Google and DeepL used in S17a, i.e., by not including the pronoun in the sentence (*Matriarchs are the rulers of the family and they usually make important decisions… > Les matriarches sont les dirigeants de la famille et prennent habituellement des décisions importantes…*). DeepL also used this solution in S17e, but Google explicitly translated the pronoun (*Les matriarches sont les dirigeants de la famille et elles prennent généralement des décisions importantes...*). Our hybrid SMT system, CNN-based system, and RNN-based systems produced similar translations as they did in S17a, where they mistranslated *rulers*, referring to *matriarchs*, as *gouvernants* or *dirigeants*, and consequently inaccurately translated *they* as *ils*. In S17f, Google and DeepL were the only systems to accurately translate *they* designating the *waitresses* as *elles*. Our hybrid SMT system and our CNN-based system struggled with agreement in the first section of the sentence (they generated *Les serveuses sont venus* instead of *venues*) and translated *they* as *ils*. Our attention-based system once again mistranslated *waitresses* as *serveurs*

and produced the same grammatically correct but inaccurate translation. Most unsusually, our RNN-based system produced a "translation" that is not only in the source language, but is also entirely unrelated to our source sentence (*The waitresses came with the bill for our table but they forgot to include some of our drinks, and the desserts that we ordered were not on the check either. > The staff were very friendly and helpful. The room was clean and comfortable*).

Our long sentences with the cue placed after the pronoun seem to be harder for the systems to translate, as S17g was the only sentence that none of the systems managed to translate accurately. Our hybrid SMT system and our CNN-based system were close to being correct as they correctly translated the pronoun *they* as *ils*, but they mistranslated the noun *nurses* as *infirmières* even though we mentioned that they are back from paternity leave. Our RNN-based system, our attention-based system, as well as Google and DeepL, all translated *nurses* as *infirmières*. Our RNN-based and attention-based systems translated *they* as *elles*, while Google and DeepL omitted the pronoun as they did in S17c. S17h was also problematic for most systems, since our attention-based system was the only system to correctly translate *they* as *elles* (*When those two joined that software engineering team of four, they turned the formerly all-male team into a mixed one that is now one-third female. > Lorsque ces deux personnes se sont jointes à l'équipe d'ingénierie logicielle de quatre personnes, elles ont transformé l'ancienne équipe entièrement masculine en une équipe mixte qui est aujourd'hui un tiers féminin.*). However, as mentioned previously, we believe this could be a fortunate use of *ces deux personnes* to designate *those two* that ultimately made the translation correct because *personnes* is feminine plural.

It was particularly hard for us to interpret the results obtained in this challenge since we could not determine whether the systems defaulted to the masculine, or if they truly were not able to make the linkages required to interpret the context. In other cases, we saw that omitting a pronoun or using gender-inclusive terminology allowed the systems to render an accurate translation while bypassing the challenge. The systems' failure to interpret words such as *matriarchs* or (male) *nurses* could also be due to their not coming across these words as often in the training data or at all in the case of our in-house systems, as our training data is restricted compared to the commercial systems. All

in all, we still found that having the cue appear before the pronoun and in the direct antecedent seems to make it easier for the systems to correctly interpret the sentence, and shorter sentences seem to yield better performance (this will be further discussed in 4.3.1.2).

4.2.2.3. Results for these

In our third challenge focusing on anaphora, we added another layer of difficulty, as the systems not only have to identify the right gender and number agreement, but also have to reflect the proximity aspect in their translations (e.g., using an equivalent such as *ceux-ci* or *celles-ci*).

	Without interruption				With interruption			
Anaphora - *these*	**18a (S)**	**18e (L)**	**18b (S)**	**18f (L)**	**18c (S)**	**18g (L)**	**18d (S)**	**18h (L)**
Hybrid	✗	✗	✓	✗	✓	✓	✓	✓
CNN	✗	✗	✓*	✗*	✗	✓	✗*	✗
RNN	✗	✗	✓*	✓*	✓	✓	✓	✗*
Attention	✗	✗	✗	✗	✗	✗*	✗	✗*
Google	✓	✓	✓	✓	✓	✗	✓	✓
DeepL	✗	✗	✓	✓	✓	✓	✓	✓

Table 32 – Results for short and long variants of sentences with the anaphora *these*

Google and DeepL did not struggle with this challenge, scoring the highest with 87.5% and 75% of the sentences accurately translated, respectively. They were followed by our hybrid SMT system and our RNN-based system, then by our CNN-based system. Surprisingly, our attention-based system, which we were expecting to produce better results, did not manage to translate any of the sentences, resulting in a score of 0%, mostly due to it not translating the proximity aspect.

For this challenge, we created short and long sentences, with and without interruption, similar to S16. We found that the systems' performance for the short and long variants is comparable, as 58.3% (or 14 of 24) of the short sentences and 45.8% (i.e., 11 of 24) of the long sentences were translated correctly. Unexpectedly, the sentences

with interruption produced more correct translations than the ones without (15 correct translations with interruptions vs. 10 without).

Google was the only system to successfully translate both of our short sentences without interruption. In S18a, we were looking for *celles-ci* as a translation for *these* referring to *pockets* (*Girls usually carry handbags because their pockets are too small, but these are very spacious.*). Most systems that failed this challenge did so because they translated *these* as *elles* and omitted to translate the proximity aspect. In contrast, DeepL failed because it translated *these* as *ceux-ci*, accurately rendering the proximity aspect, but incorrectly identifying the noun to which *these* refers, possibly associating *these* with *handbags* rather than *pockets* (*Les filles portent généralement des sacs à main parce que leurs poches sont trop petites, mais ceux-ci sont très spacieux.*). S18b was seemingly easier for systems to translate, as our attention-based system was the only system to incorrectly translate the sentence, failing to render the proximity aspect (*engagement rings, and these are always over-priced > anneaux d'engagements, et ils sont toujours surévalués*). Our hybrid SMT system, Google, and DeepL all found the same solution, which was the solution we were expecting (*celles-ci* for *these* referring to *engagement rings*, i.e., *bagues de fiançailles*). Our CNN-based system and our RNN-based system went another route: Our CNN-based system used *qui* (*engagement rings, and these are always over-priced > anneaux d'engagement, qui sont toujours surévalués*). *Qui* was not a solution that we were considering, but it is nonetheless correct as it satisfies our three criteria for assessment. Our RNN-based system used *ceux-ci*, likely because it unexpectedly translated *engagement rings* as the masculine noun phrase *cercles d'engagement* (*cercles d'engagement, et ceux-ci sont toujours sur-évalués*); the translation was thus correct.

Our short sentences with interruption produced results that were a little more consistent, with our CNN-based system and our attention-based system being the only two systems to fail to accurately translate the challenging element in both sentences. In S18c, all the other systems managed to accurately translate *these* referring to *cottage owners* as ceux-ci, but our CNN-based and RNN-based systems translated *these* as *ils*, omitting the proximity aspect. In S18d, our CNN-based system produced an odd translation where it completely ignored the challenging element and added a new,

unrelated sentence instead (*I thought I owned the ugliest shoes in three counties, but these are even uglier > J'ai pensé que j'étais propriétaire des chaussures uglies dans trois comtés, mais il s'agit même d'un bout à l'autre. Il s'agit d'un problème de santé*). Our attention-based system failed the challenge because, once again, it did not render the proximity aspect (*J'ai pensé posséder les chaussures les plus uglies dans trois comtés, mais elles sont même plus aberrantes*). Our hybrid SMT system, Google, and DeepL all produced the same correct translation (*mais celles-ci sont encore plus laides*). Our RNN-based system, on the other hand, found a solution that involved repeating the noun to add emphasis (*J'ai pensé que j'avais les plus beaux chaussures dans trois comtés, mais ces chaussures sont même ugelles*). This translation is correct because there is no need for agreement: the system repeated the noun, therefore proving it identified the right item. Furthermore, the challenging element does not require post-editing.

There were more incorrect translations in our long sentences without interruption than in our short sentences without interruption. Results for S18e were identical to the results for its short variant S18a: Google was the only system to accurately translate the challenge using *celles-ci*, while all the other systems failed because they omitted to translate the proximity aspect (they translated *these* as *elles*), with the exception of DeepL, which failed because it had the wrong gender agreement (*ceux-ci* instead of *celles-ci*). Results for S18f, however, showed that the long variant was noticeably harder for systems to translate, with half of the systems failing the challenge, as opposed to only one in the short variant. In S18f, our hybrid SMT system inaccurately translated *these* referring to *wedding bands and engagement rings* as *ceux-ci* (*wedding bands and engagement rings, and these are always over-priced > des bandes de mariage et de fiançailles, et ceux-ci sont toujours trop chers*). Our CNN-based system ignored the challenging element and added a phrase (starting with a capital letter) relating to a hotel location (*L'hôtel est très bien situé, à proximité de l'aéroport*), a pattern we found in the past with our RNN-based and attention-based systems in S5d, with our CNN-based and attention-based systems in S5h, as well as with our CNN-based system in S15g. As for the third system to fail the challenge, i.e., our attention-based system, it generated the same incorrect translation as it did in the short variant by omitting to include the proximity aspect in its translation. Google and DeepL successfully translated the

challenging element as *celles-ci* as we were expecting. Our RNN-based system, however, translated *these* as *ceux-ci*, but that translation was marked as correct nevertheless, as the system chose to translate *engagement rings* as the masculine *anneaux d'engagement*, instead of the feminine *bagues de fiançailles* that we had in mind.

Our hybrid SMT system and DeepL were the only systems to accurately translate both long sentences with interruption. In S18g, our hybrid SMT system, our CNN-based system, our RNN-based system, and DeepL all correctly translated the challenging element as *ceux-ci*. Google translated *these* as *ils* and, therefore, omitted to translate the proximity aspect, whereas our attention-based system ignored the challenging element (*Most cottage owners are residents of nearby cities […], but these came from England > La plupart des propriétaires de chalets sont des résidents de villes voisines […], mais viennent d'Angleterre*). In S18h, aside from our hybrid SMT system and DeepL, Google also managed to translate *these* referring to *the ugliest shoes* accurately as *celles-ci*, which stands for *les chaussures les plus laides*. Our CNN-based system added a phrase that may be somewhat related at the end of the sentence instead of translating the challenging element (*but these are even uglier > mais Il s'agit même d'une horreur*) and omitted to translate *these* in this sentence; hence, the translation was marked as incorrect. Our RNN-based system omitted to translate the end of the sentence, which included the challenging element (*with the weird lace colour, ridiculous platform, and tacky pattern, but these are even uglier > avec la couleur des dentelles bizarres, la plate-forme ridicule et les motifs.*). As for our attention-based system, instead of explicitly translating the challenging element, it generated the following translation: *J'ai pensé posséder les chaussures les plus uglies dans trois comtés, avec la couleur de lace, une plate-forme ridicule, et des motifs aberrants, mais même aberrants*, where it seems to have interpreted *but these are even uglier* as *but even uglier*.

All in all, we found that most systems that failed to translate the challenge failed because they omitted to take into account the proximity aspect. This was not unexpected, given that it was an additional challenge added to an already demanding task. We also found, despite the fact that our attention-based system failed to accurately translate all eight sentences, that it did better in the short sentences and got worse and worse as the sentences got longer and as interruptions were added. In the short sentences, though it

failed to render the proximity aspect, it still managed to identify the correct agreements. In the long sentences without interruption, it performed similarly to how it did in the short sentences. However, in the long sentences with interruption, it started to avoid the challenging element (we will further discuss length and complexity in 4.3.1.1).

4.3. Key findings

Through our challenge set, we were able to identify some recurring patterns in the systems' output and what we assume could be default choices/strategies for dealing with specific phenomena or items. In this section, we will be providing a summary of and deeper dive into our previous findings, as well as presenting additional findings, as we noticed some patterns in other parts of the sentence that are worth analyzing. We divided our key findings into two sections each, first looking at patterns found through all systems, then going into patterns found in specific models. Similarly, our section on additional findings will be divided into difficulties affecting all systems and difficulties affecting specific systems.

Finally, we will review the sentences that were the most and least successfully translated by all systems and we will provide either advice on what makes a good sentence for the challenge set, or possible modifications we could make to our sentences to improve them for possible future use.

4.3.1. General key findings

In analyzing the results of our challenge set, we found that certain challenges and certain variants were more problematic for systems to translate than others. In this section, we will review challenges and variants that all systems struggled to translate or all successfully translated, and try to interpret the results.

Overall, of the total of 144 sentences in the challenge set, there were ten that all the systems managed to successfully translate, seven of which were short variants. Of these ten sentences, six were the short and long variants of the same sentence, and four were of individual short sentences, as shown in Table 33. The rarity of universally correctly translated sentences suggests that the challenge set—as expected based on our initial testing with larger-scale online NMT systems—targeted phenomena that were

difficult for at least one of the systems. This is evidence of that the improvements of NMT performance have still not overcome the usefulness of a challenge set approach for studying potential pitfalls, and the approach remains useful for studying difficult problems.

Challenge	Short only	Short and long pair
As (cause)	S2c, S2d	
While (concession)		S5b-S5f
While (opposition)	S6c	S6d-S6h
With (feeling/state)	S11d	
Anaphora (*it*)		S16d-S16h

Table 33 – Sentences that all systems successfully translated

The fact that long variants were less likely to be universally correctly handled does help to support the hypothesis that sentence length correlates unfavourably with system performance. This hypothesis will be investigated in more detail in section 4.3.1.1. However, the correlation between short and long variant performance indicates that not only sentence length, but also the specific problems themselves affect overall outcomes. Unfortunately, however, no recurrent characteristics typical of sentences that were universally successfully translated could be identified in the results, so we are unable to make any judgments in that respect.

Challenge	Short only	Long only	Short and long pair
While (temporality)		S4g	
While (concession)		S5h	
When (causality)	All of S7		
When (continuity)	All of S8		
When (in spite of)	S9a		S9c-S9g
With (causality)	S10c		S10a-S10e; S10d-S10h
With (in spite of)	All of S12		
Homographs		S13g	S13a-S13e
Anaphora (*they*)		S17g	

Table 34 – Sentences that all systems failed to translate accurately

Various subsets of our findings can be analyzed through a number of lenses: length of sentences, position of the challenging element, and effect of interruptions.

These groups represent potential reasons why the systems might "struggle" to translate a challenging element.

4.3.1.1. Length of sentence

Overall, the results (summarized in Table 35) showed that sentence length does not seem to affect the systems' performance as much as we thought. As discussed in section 3.1.1, we hypothesized that—as often described in the literature—longer sentences would tend to pose greater challenges than shorter ones, regardless of the specific type of challenge the sets targeted, in large part due to the complexity of long-range dependencies. In half of the challenges, the overall results for all experimental systems (i.e., our hybrid SMT, CNN-based, RNN-based, and attention-based systems) were the same for the short and long variants. In the other half, we noted seven challenges where the short variants yielded better results than their long counterparts (highlighted in green in Table 35), and only two challenges where it was the opposite case (highlighted in red).

Evaluating at a more detailed level how each system performed when faced with different lengths of sentences was tricky, as different challenges produced different results. We do, however, recognize that certain challenges might be harder to translate than others when we take length into consideration (some of these sentences will be further discussed in 4.3.1.2).

For example, in sentences where we are testing semantic ambiguity in the source language (S1 to S12), we are hypothesizing that the systems will rely on semantic cues found elsewhere in the sentence to help inform disambiguation of lexical items, such as *as*, *while*, *when*, and *with*. In theory, longer sentences would yield better results, or at least just as good, since the systems have access to more data points. We found this to be true to some degree, as the results were comparable in the short and long variants (or higher in the long sentences the case of S9) in the majority of our challenges resulting from semantic ambiguity in the source language.

In contrast, challenges such as anaphora (S16 to S18) require the systems to explicitly link two items that likely appear farther apart the longer a sentence becomes, potentially making the task more complex. In those cases, we did notice that two of the

three challenges involving anaphora produced better results in the short variants than in the long ones.

Challenge	Short	Long	All
S1: “as” expressing simultaneity	43.8%	43.8%	43.8%
S2: “as” expressing a cause	87.5%	68.8%	78.1%
S3: “as” expressing progression	43.8%	43.8%	43.8%
S4: “while” expressing temporality	68.8%	25.0%	46.9%
S5: “while” expressing a concession	56.3%	56.3%	56.3%
S6: “while” expressing an opposition	62.5%	62.5%	62.5%
S7: “when” expressing causality	0.0%	0.0%	0.0%
S8: “when” expressing continuity	0.0%	0.0%	0.0%
S9: “when” meaning “in spite of the fact that”	6.3%	12.5%	9.4%
S10: “with” expressing causality	12.5%	12.5%	12.5%
S11: “with” expressing a particular feeling or physical state	43.8%	31.3%	37.5%
S12: “with” meaning “in spite of”	0.0%	0.0%	0.0%
S13: Homographs	18.8%	6.3%	12.5%
S14: Scope of modifiers	18.8%	12.5%	15.6%
S15: Scope of conjunctions	25.0%	37.5%	31.3%
S16: “it”	50.0%	50.0%	50.0%
S17: “they”	25.0%	12.5%	18.8%
S18: “these”	43.8%	31.3%	37.5%
All	33.7%	28.1%	30.9%

Table 35 – Percentage of correct translations of short and long sentence variants, all experimental systems

Furthermore, when looking at how length of sentences might affect a system’s performance, we also had to consider cases where all the elements required for a system to process the challenging element were close together even in long sentences. This was the case for the sentences that we created to test for scope of modifiers (S14) and scope of conjunctions (S15). Results varied for the two challenges, as we observed more correct translations in the short variants of S14, but more correct translations in the long variants of S15. This leads us to believe that sentence length might not have as big of an influence on the systems’ performance in most cases.

In 4.4.2, we will discuss some of the specific cases where the systems seem to produce better results in the long variants than in the short ones. While these cases did

not affect how the systems translated the challenging element itself, it does, nonetheless, give us insight as to how the models are functioning when they are processing a sentence.

4.3.1.2. Position of the challenging element

As discussed in Section 3.1.2, when testing ambiguous lexical items (i.e., *as, while, when, with*) that could be placed at the beginning of a sentence or near the middle, between items they linked, we included equal numbers of both structures in the challenge set, for two reasons. First, this better reflects the potential for overall system performance on those items however they may appear in texts to be translated, and second, it allows for some preliminary comparisons of performance between the two kinds of structures. We found this particularly relevant to explore, as long-range dependency is a recognized challenge for MT systems. Although it is hard to look at the positioning of the challenging element without considering sentence length, we tried to focus our observations on how the position of a linking word can influence a system's analysis of its meaning. Namely, we are interested in seeing if a system's architecture (and therefore how it processes a sentence) might have an effect on its choice of translation. In RNN-based systems, for example, sentences are processed in a sequential manner (Yarats *et al*., 2017). As a consequence, an ambiguous lexical item placed at the beginning of a sentence might be harder to translate than an ambiguous lexical item placed between two items it links.

Table 36 below compares the percentage of correct translations of the ambiguous lexical items we tested in different positions. (Note that S9 and S11 could only occur in one position and therefore have been excluded from the table, although the data is discussed below as a basis for comparison). Overall, we found that sentences where the challenging element was placed mid-sentence yielded very slightly better results than sentences where the challenging element was placed at the beginning of the sentence. However, there were more challenges where the results for the individual sentences with the challenging element at the start of the challenge were higher than those with the challenging element placed mid-sentence.

Of the ten challenges, there were five with more correct translations when the challenging element was placed at the beginning (highlighted in green in Table 36) and

only two when the challenging element was placed mid-sentence (highlighted in red). The other three challenges produced translations that were all wrong, regardless of the positioning of the challenging element.

Challenge	Before	Mid	All
S1: “as” expressing simultaneity	50.0%	37.5%	43.8%
S2: “as” expressing a cause	75.0%	81.3%	78.1%
S3: “as” expressing progression	56.3%	31.3%	43.8%
S4: “while” expressing temporality	62.5%	31.3%	46.9%
S5: “while” expressing a concession	75.0%	37.5%	56.3%
S6: “while” expressing an opposition	31.3%	93.8%	62.5%
S7: “when” expressing causality	0.0%	0.0%	0.0%
S8: “when” expressing continuity	0.0%	0.0%	0.0%
S10: “with” expressing causality	25.0%	0.0%	12.5%
S12: “with” meaning “in spite of”	0.0%	0.0%	0.0%
All	37.5%	38.8%	38.1%

Table 36 – Percentage of correct translations of challenge sentences by challenge item placement, all experimental systems

At a more detailed level, it was hard for us to establish whether the positioning of the challenging element really affected how the systems performed or not because of the varied results we obtained.

For example, in S6 (*while* expressing an opposition), it seemed as though sentences where *while* was placed between the propositions (e.g., *The blue mushroom shrinks Mario, <u>while</u> the red one makes him grow*), regardless of their length, were correctly translated more often than sentences where *while* was placed before the propositions (e.g., *<u>While</u> Megan loves a good steak, her brother doesn’t eat meat*) – contrary to the other two meaning of *while*, which had more correct translations in sentences where *while* was placed either before the propositions or at the beginning of the sentence. This could perhaps be because some senses might be more likely to occur in certain positions in a sentence and the system might be able to recognize that. It would be worth exploring in future work, as it could potentially help tweaking of MT systems (e.g., with rules), but also inform writing for translation.

It was also difficult for us to draw any conclusions on positioning when we had poor results overall. For instance, in S8 (*when* expressing continuity), we found that the

position of *when* does not seem to affect the system's performance, as all the systems used *lorsque*, *alors que*, and *quand* interchangeably. However, due to all the systems failing the challenge, we were not able to determine if a certain structure truly leads to a certain solution.

4.3.1.3. Effect of interruptions

We tested the effect of interruption in two of our challenges involving anaphora. As mentioned in Section 3.4.2, we decided to add interruption in our sentences to add a layer of difficulty to the challenge. Anaphora resolution has been known to be a challenge in MT (Arnold, 2003, p. 119), and although we thought that adding interruptions would make it harder for systems to translate our sentences, we were surprised to see that, overall, results for sentences with interruption were equal to or higher than those for sentences without interruption (summarized in Table 37). While recognizing that our sample is too small to truly reflect the systems' overall handling of anaphora, we were nonetheless able to observe that more interruptions seems to make it harder for systems to accurately translate the sentence.

Challenge	Without interruption	With interruption	All
S16: "it"	43.8%	43.8%	43.8%
S18: "these"	25.0%	50.0%	37.5%
All	34.4%	46.9%	40.6%

Table 37 – Percentage of correct translations of sentences by challenge with and without interruption, all experimental systems

In S16 (anaphora – *it*), for example, we found that increasing the distance between the noun and the pronoun (i.e., by adding more words) did not affect how the systems performed as much as adding more interruptions (i.e., nouns). Some of our sentences with a greater number of words between the noun and the pronoun achieved higher results that our other sentences where there were fewer words between the two elements, but more nouns among those words (see discussion of S16c and S16d in 4.2.2.1). This suggests that ambiguity poses a greater challenge to the systems than sentence length.

We also found that, in sentences with *they* (S17), having the gender cue appear before the pronoun and in the direct antecedent seems to make it easier for the system to interpret the sentence than having it after, or in a word other than the direct antecedent. For instance, in our sentence S17a, "The *matriarchs* are the rulers of the family and *they* usually make important decisions", we believe that including the plural noun "rulers" made the sentence more ambiguous for the systems, whereas in our sentence S17b, "The *waitresses* came with the bill, but *they* forgot to include some of our drinks", with no other plural noun included, it seems to have made it slightly easier for systems to translate.

4.3.2. Other elements affecting performance

In addition to the elements such as length, positioning of the challenging element, and interruption that we intentionally included in our challenge set for the purpose of analyzing their effect, we also noticed other elements that might have influenced the systems' performance and that go beyond difficulties that we were anticipating. These elements include ranking of the senses for ambiguous lexical items, possible default translations, effect of ambiguity, handling of subtle differences in meaning, and effect of training data. In the next sections, we will present them and try to interpret their effects.

4.3.2.1. Ranking of senses for ambiguous lexical items

In some of our sentences testing lexical ambiguities, we found that systems failed to translate some meanings more often than others. In S2, for example, the systems struggled to translate the causal *as* more than they did translating S1, *as* expressing simultaneity. In searching for an explanation for this phenomenon, we wondered whether the frequency of the various senses might affect the likelihood of their being translated appropriately. This is a plausible scenario, since, as discussed in 1.1.5, NMT systems "learn" translations by analyzing corpora of translated texts, and (like SMT systems before them) make choices based largely on probabilities of the most likely translation according to the data available. To help us evaluate this possibility, we turned to the LDOCE, as it ranks its definitions by showing the most common meanings of a word first. In Table 38 below, we compare the ranking of the senses we tested in the LDOCE

with the systems' performance for each of the senses, to attempt to determine if there may be any correlation.

Challenge	LDOCE sense number	Overall system performance
S1: "as" expressing simultaneity	4	43.8%
S2: "as" expressing a cause	5	78.1%
S3: "as" expressing progression	4	43.8%
S4: "while" expressing temporality	1/2	46.9%
S5: "while" expressing a concession	4	56.3%
S6: "while" expressing an opposition	3	62.5%
S7: "when" expressing causality		0.0%
S8: "when" expressing continuity	3	0.0%
S9: "when" meaning "in spite of the fact that"	8	9.4%
S10: "with" expressing causality	14	12.5%
S11: "with" expressing a particular feeling or physical state	4	37.5%
S12: "with" meaning "in spite of"	19	0.0%
All		32.6%

Table 38 – Percentage of correct translations of challenge sentences by frequency-based sense ranking in the LDOCE, all experimental systems

While we established that a lower ranking in terms of usage frequency can be associated with with lower-quality results from the systems, there were cases where it was hard to determine whether lower frequency of the sense was truly the *cause* of the systems failing or not. In S12 (*with* meaning "in spite of"), where the systems mistranslated all the sentences, we were not able to determine whether the systems failed because the link expressed by *with* is more abstract (and therefore it is harder for the system to identify cues), if it is because of the lower frequency (and consequently the right meaning is lower-ranked in the LDOCE entry), or if it is a combination of the two.

Nevertheless, we believe that the variations in frequency can play a role in how the systems translate the lexical unit in a given sense, especially if the training data is representative of the ranking (meaning that senses that are less commonly used are also less commonly found in the training data).

4.3.2.2. Possible default translations

In many cases, and especially in our sentences with ambiguities in the source language, we think some systems could have resorted to a default translation.

In S6 (*while* expressing an opposition), all the failures were due to the systems using *bien que* and *même si* (both of which would express concession rather than opposition). In S7 (*when* expressing causality), all the translations deemed inaccurate were because the systems translated *when* as *lorsque* (instead of *parce que*). (It should nevertheless be noted that this sense is a bit tricky for a machine, because it would be required to make explicit in the TL what is implicit in the SL). In S9 (*when* meaning "in spite of the fact that"), all the systems that accurately translated the challenge used *alors que* as their solution. This seems to be a safe solution that works in many contexts and for many polysemous words, as seen in S1 (*as* expressing simultaneity) and S6 (*while* expressing an opposition).

When we looked at deixes, for S17 in particular (*they*), we were anticipating the systems to default to the masculine (*When those two joined that team, <u>they</u> turned the all-male team into a mixed one*), but it was particularly hard for us to interpret the results obtained in this challenge because we could not determine whether the systems truly defaulted to the masculine, or rather if they were simply not able to make the linkages required to interpret the context. In other cases, the systems managed to bypass the challenge either by omitting the use of a pronoun or by using gender-inclusive terminology. (For example, our CNN-based system's translation of *nurses* in S17c was *infirmières et infirmiers*.)

While we were not able to determine if all systems had a default translation for a particular challenge, we were able to find certain patterns in specific systems that could pass as a potential "go-to" translation. These will be discussed in 4.3.3.

4.3.2.3. Effect of ambiguity

Though we did our best to create clear sentences, there were some occasions where we unintentionally included language that the systems found ambiguous, mostly due to part-of-speech ambiguity. For example, in S3d (*To make risotto, you must add small*

quantities of broth as the rice cooks), all of our systems struggled to translate *cooks* as a verb and instead translated it as a noun. This may have caused them to fail to translate the challenging element, as they interpreted *as* expressing progression as *as* used for comparison. This also happened in in the long variant, although in S3h our attention-based system was successful in translating *cooks* as a verb.

We observed a similar phenomenon in S9g (*I was upset and our whole friend group was frustrated that she still trusted and dated him, when I warned her that he was a cheater*), where all of our systems mistranslated *dated* as *datée/daté* instead of, e.g., *soit sortie avec lui*. In this case, we were able to determine that it was probably because the meaning "to write or print the date on something" is more commonly used than the meaning "to have a romantic relationship with someone/go out with", as ranked in the LDOCE.[45]

Seeing that most of our experimental systems (which have limited training data) failed to translate these sentences accurately, but that Google and DeepL did not find these parts of speech challenging, we can argue that more training data might help eliminate these misinterpretations and allow the systems to build a stronger context to pass the challenge.

We also found another type of phenomenon involving part-of-speech ambiguity, where we believe the systems mistranslated a word, not because of a lack of training data, but because it was being used in an unusual way. In S13c and S13g, we used the word *like* as derived noun, and most systems interpreted it as meaning *comme* (which makes DeepL's translation of the sentence even more impressive in S13c). In cases like these (and for homographs in general), we believe human post-editing may almost always be needed.

4.3.2.4. Handling of subtle difference in meaning

While the goal in creating the challenge set was to use lexical units and meanings that could reasonably be distinguished by both humans and MT systems, and would not be too specialized or too rare to be known to the MT systems, it was still difficult or

[45] The former of these two senses is listed first in the LDOCE entry for the verb, while the latter is the fourth sense in the entry.

impossible to ensure this was criterion was met, in part because we could not easily establish the number of occurrences of a given lexical unit (much less a particular sense of it) in the training data. It was thus difficult to determine how common a given sense was in the data, and rare units or senses might be expected to correlate with poorer MT performance.

In S8 (*when* expressing continuity) for example, the sense is ranked third of nine in the LDOCE entry for *when*, a rank we considered was not too low for the sense to not be frequently found in usage. However all the systems translated *when* as expressing simultaneity (first and second sense) or simultaneity and opposition, but not continuity.

In S18 (anaphora – *these*), most of the systems that failed to translate the challenge failed because they did not take into account the proximity aspect. In creating this challenge, we were trying to add a layer of complexity to the analysis: having looked at the temporal aspect expressed by challenging elements in previous challenges, we now also wanted to look at the spatial aspect. This was hard for us to achieve as we were restricted to a sentence-level analysis and as we were also restricted in terms of number of words per sentence. We identified *these* as being a possible candidate, although we realize that the proximity aspect expressed by *these* is mostly found in spoken phrases (the LDOCE entry for *these* refers back to the entry for *this*, and in the entry for *this*, the meaning that we were looking for is the fourth sense, but is identified as being a "spoken phrase").

We recognize that some senses often come with explicit cues that might help systems identify the correct meaning. (e.g., temporal senses could be triggered by vocabulary about time of day or year, or combination of verb tenses used to express interruption), while others, such as S18, may not have those cues. This is difficult to quantify and although it would be interesting to do a more fine-grained analysis, it is beyond the scope of this project. It would be worth exploring, however, in future work, in order to provide some insight into the likelihood of MT systems being able to handle subtle difference in meaning using context.

4.3.2.5. Effect of training data

Although the focus of this thesis was not to determine how training data could influence a system's performance, there were some challenges where we believe the amount and/or type of training data could have played a direct role in the translations produced.

The sentences we created for S11 (*with* expressing a particular feeling or physical state) were evidently easier for Google and DeepL to translate than for our in-house systems. On the one hand, we think this could be because the challenging elements in this challenge were close to being collocations, thus, not involving processing so much as retrieval by the systems. On the other hand, in these expressions, *with* was associated with words that are often linked to emotions (e.g., *hatred*, *joy*), which are more likely to be found in Google and DeepL's training data, through data collected from usage. As such, we believe this may have contributed to the differences in system performance, specifically the poorer performance of the experimental systems. We also think that adding to our systems' training data could alleviate the problem, as they did not seem to struggle with the sentence in which *with* was associated with a medical condition rather than an emotional state (*That medication is not recommended for use in patients with hypothermia*), which is more likely to be similar to the content of our training data. We used data that had been crawled from the web or that consisted of official proceedings or of government websites material (for more detailed information, refer to Section 2.2.1), whereas data obtained from users on Google and DeepL might also include translation of personal correspondence, which are more likely to be emotion-rich.

Increasing the amount of our training data could also help with words that we did not think were "rare" when we created our challenge set. For example, in S14d and S14h, all of our NMT systems failed to translate *crisp* accurately, instead translating it as *criant*, *cristalline*, *critiquée*, or even *cristé*. Our hybrid SMT system, however, managed to translate it as *croquante*, so we believe the word was found in the training data, but perhaps did not have the weight needed (i.e., was not found frequently enough) to be considered as a likely translation by our NMT systems.

4.3.3. Key findings in specific models

While our challenge set did not reveal a large number of key findings that applied to all systems, it did highlight many patterns that seem to be specific to certain systems. In the following sections, we will elaborate on these patterns and try to link the solutions the systems came up with to the models' architecture.

We were able to identify more patterns in our CNN-based system's translations than in our other two systems, as most of our findings originated with difficulties that the systems encountered. Therefore, the system that struggled to translate our challenges most often was also the system that revealed the most about its functioning. To complement our analysis, we also looked at patterns recurring in our hybrid SMT system's translations, as they were useful for spotting differences between our NMT systems and what was considered previously as being "state-of-the-art".

4.3.3.1. Hybrid SMT

Because, as we described in Section 1.1.4.1, SMT systems work by generating all possible TL sentences and finding the most probable one, we expected our hybrid SMT system might have "default" translations for some polysemous words. Some of our findings did indeed support this hypothesis: in S1 (*as* expressing simultaneity), for example, we found that the system almost always defaulted to *comme* as a translation for *as* expressing simultaneity. While this caused the system to be systematically wrong in S1, it also allowed it to be correct mostly correct in S2 (*as* expressing a cause). We hypothesized that the system could have a default translation for all senses of *as* and this could be partially true, as the only case where it used another solution was in S2g (*as the young patient had experienced discomfort > en tant que jeune patient avait connu l'inconfort*). In most of our other sentences, we had a structure that consisted of *as I* or *as* followed by a clause that sometimes has a non-human subject, but in S2c and S2g, we have a structure that closely resemble the second sense listed ("used to say what job, duty, use, or appearance someone or something has") in the LDOCE entry for the preposition *as*. This ruled out the universal "default" translation hypothesis—which, in

reality, would be more typical of a rule-based model—but still supports the idea that the system might have a "favoured" translation.

Another hypothesis is that, in cases where the sense distinctions are relatively fine and there are no very obvious cues in the immediate context to assist in disambiguating, our hybrid SMT system is simply picking the most statistically probable translation candidate as its solution, and that more frequently used senses may be found more often in the training data, making them more likely to be "learned" and subsequently called on by our system. The findings would be consistent with such a hypothesis, since *as* used to compare two things is the first sense in the LDOCE entry, indicating highest frequency, and would also be translated as *comme*.

Similarly, in S4, our hybrid SMT system failed to translate *while* expressing temporality, using *alors que* and *tandis que* in all of its sentences. *While* meaning "during the time that something is happening" is the first sense in the LDOCE entry, and would translate to *alors que* or *tandis que*, whereas the meaning we were looking for, "all the time that something is happening," is the second sense. This again supports our new hypothesis. These are all observations that we will be able to use as a basis for comparison with our NMT systems, to see if they show similar patterns or if, as expected from the characteristics described in literature, they will rely more on context analysis to choose the most likely equivalent.

In S3 (*as* expressing progression), we included an incomplete comparison (*comparatif elliptique*) in S3c (*We will move to a bigger house as our family grows*). Our hybrid SMT system translated *as* as *que* in this sentence, suggesting a smaller window of context where the system possibly analyzed the phrase *bigger house as our family* first, without the verb *grows*. This could give the impression that *as* is used for comparison, rather than to express progression, although the structure is not quite typical. While our RNN-based and attention-based systems manage to accurately translate this sentence, our CNN-based system came up with the same incorrect solution as our hybrid SMT system, suggesting that it might also work with a smaller window. This is coherent with our knowledge of the architecture, as CNN-based system focus on particular sections of a sentence at a time and may not always consider the sentence as a whole (see 1.1.5.1.2).

In S11 (*with* expressing a particular feeling or physical state), the system produced different translations for the short and long variants of a sentence (S11d and S11h, both of which were correctly translated). In S11d, the system translated *patients with hypothermia* as *patients atteints d'hypothermie*, and in S11h, it translated *patients with thyroid problems* as *patients présentant des problèmes thyroïdiens*. This complements our previous hypothesis, where we stated that our hybrid SMT system may have a smaller window of context, and this example further shows it may be looking at only a few words immediately before and (especially here) after. By inserting a new item between *with* and its original cooccurrent (*with hypothermia > with thyroid problems or severe hypothermia*), we probably changed how our hybrid SMT system processed the chunk containing the challenging element.

This limited context window has proven to be advantageous in other challenges. For instance, we found that our hybrid SMT system was unexpectedly stronger than our in-house NMT systems at translating scope-related challenges. As mentioned in Section 4.2.1, SMT models partly rely on the frequency of a translation in their training data to determine the most likely translation candidate. As such, their context window might be limited, but if the phrase they need to translate can be found in the training data, it is easier for an SMT system to simply retrieve it, than for an NMT system to build a context around it to reach the most likely translation. In other words, instead of having to analyze an entire sentence in order to identify the appropriate scope, the system can rely on its limited window of context and potentially retrieve parts of the challenge that may have appeared before in the training data.

While it allowed our hybrid SMT system to successfully translate sentences with scope difficulties, the limited window of context may also have been a disadvantage in some of our sentences with anaphora. The system found S17 (*they*) particularly hard to translate, and we think this could, in part, be because the cues we included in the sentences were perhaps too distant from the pronoun, i.e., excluded from the chunk including *they* and, therefore, processed separately. This challenge was the only challenge among the set of anaphora challenges that relied on the location of a cue (*The matriarchs are the rulers of the family and they usually make important decisions*), rather than interruptions in the sentences (*The recital was tonight but there were issues with the stage*

so it was cancelled), and this was the only one of the anaphora challenges that our hybrid SMT system failed to translate entirely. However, our NMT systems also struggled with this challenge the most, among the anaphora challenges, and we believe it may be because, as explained in 4.2.2.2, sometimes the cue was not in the antecedent itself and locating it is harder than identifying the correct antecedent.

4.3.3.2. CNN-based

We found more unexpected solutions from our CNN-based system than our hybrid SMT system, although we were not necessarily surprised to see these solutions. As stated in 1.1.5, NMT systems are known to be "black boxes" and are more likely to produce unexpected translations (including hallucinations). In one example of unpredictable behaviour, there were several occasions where it omitted to translate the challenge entirely. For example, in S1 (*as* expressing simultaneity), it sometimes ignored the challenge (e.g., S1b: *As I put my curtains up, I realized the rod was crooked > J' ai mis mes rideaux, j' ai réalisé que la tige était tombée*) and sometimes translated the simultaneous *as* as a causal *as* instead (e.g., S1c: *My knees cracked as I stood up > Mes genoux se sont fissurés comme je l' étais*). In S6g (*while* expressing an opposition), the system not only omitted the challenging element, but also started a new sentence (*I lost faith in humanity when I saw that there was still hand soap on the shelves, while the toilet paper was out of stock > J' ai perdu confiance en l' humanité quand j' ai vu qu' il y avait encore du savon à main sur les étagères. Le papier de toilette n' était pas en stock*). Then, in S9c (*when* meaning "in spite of the fact that"), it omitted the challenging element and also reworded the sentence, shifting its meaning entirely (*I was upset she still dated him when I warned her he was a cheater > J' ai été bouleversée qu' elle l' avait toujours avertie qu' il était un tricheur*). While the sentence produced is grammatically correct, the translation is inaccurate and would require post-editing.

The sentence splitting that our CNN-based system seems to occasionally do has some variants. In S6g, the system ended the sentence with a period and started a new one with a capital. In S2h (*as* expressing a cause), it produced a structure that is characteristic of French (*comme suit:*) but does not make sense in the sentence. A similar structure was also found in S3g (*as* expressing progression), where it translated "*as* our family grows"

as "*à savoir:* Notre famille grandit". The system seems to use this structure that includes a colon in several sentences, as we, again, found it in S14d (scope of modifiers), where it added a colon mid-sentence as to introduce an example or explanation (*listen to my abnormally crisp asparagus story > écoutez mes anormalement criants: L'histoire de l'épereur*). These emerging "sentence-splitting" patterns are evidence that the structure of the sentence may play a considerable role in the handling of challenges for certain systems.

4.3.3.3. RNN-based

Not many patterns emerged from our analysis of how our RNN-based system translated the challenge elements in our sentences. Most of our findings for this model were found elsewhere in the sentence and will thus be discussed in 4.4.2.

We did notice, however, that in S3 (*as* expressing progression), our RNN-based system failed to translate the word *cooks* accurately in the short variant (S3d: *as the rice cooks > comme les cuisiniers*), but managed to translate it correctly in the long variant (S3h: *lorsque le riz cuira*). This could suggest that our RNN-based system works with a larger context of window and performs better when provided with a larger context, as was the case in the long variant. This is consistent with observations from the literature, as RNN-based systems work in a sequential manner and might perform better the more context they have to work with. In this case, we added more context not only before *cooks*, but also immediately after it (*you must add small quantities of broth <u>as</u> the rice cooks > you must add the chicken or vegetable broth in small quantities <u>as</u> the rice cooks over medium heat*). Although it does not touch upon the challenging element that we were trying to test, this correct translation of the word *cooks* may have influenced the translation of *as*. The reason for this is, as mentioned in 1.1.5.1.1, the first RNN in an RNN-based system converts each unit of information that it reads sequentially into a list of unique measurements, and the accuracy of these measurements depends on the context provided.

4.3.3.4. Attention-based

The difficulties our attention-based system had seem to either be related to length or ambiguity. In S1 (*as* expressing simultaneity), the system accurately translated most sentences, but failed to translate S1c and its long variant S1g. In S1g in particular, we think the system struggled to translate the sentence because of the multiple temporal conjuctions found in it (*My grandmother laughed at me and told me I was getting old when my knees cracked as I stood up to go grab my phone*). It seems to have made it harder for the system to build a context: the system managed to translate the first temporal conjunction (*when > lorsque*) but omitted to translate the second one (*my knees cracked as I stood up to go grab my phone > mes genoux se sont fissurés pour prendre mon téléphone*). Conversely, it seems to have struggled to translate the short variant S1c because the sentence may have been too short (seven words) for the system to build an accurate enough context.

This was a surprising case, as in other challenges, sentence length seems to affect our attention-based system's performance overall. In S18 (anaphora – *these*), despite our system failing to translate all eight sentences, we noticed that it still did better in the short variants and got worse and worse as the sentences got longer and as interruptions were added. As the sentences became harder to translate, our attention-based system started to avoid the challenge and omitted to translate the challenging element in our long sentences with interruption.

The system also found ambiguous verbs hard to translate. For example, in S2e (*as* expressing a cause), we used *kept being* to express repetition rather than continuity (*As I kept being woken up by raccoons opening my garbage bins every night, I put traps with peanut butter and jelly in my backyard.*). Instead, the system interpreted *kept being* as a continuous action that is being interrupted by the next action (*I put traps*) and translated *as* as *alors que* consequently.

4.3.3.5. Google

Though we only meant to use Google and DeepL as "benchmarking" systems that are more representative of the recent NMT models and the volume of training data

potentially available than our in-house systems, we also noted certain patterns in each system that are within the scope of our challenge set.

While we were analyzing Google's translations of the sentences with *as*, we found that in S1 (*as* expressing simultaneity) and S2 (*as* expressing a cause), Google often used *alors que* as its translation (with the exception of three sentence pairs: S1b-S1f, S2c-S2g, and S2d-S2h). At first, we did not know if it was potentially a default translation for *as*, but after seeing more diversity in S3 (use of gerunds and of solutions such as *à mesure que*, *au fur et à mesure que*), we deduced that the difference between *as* expressing simultaneity and *as* expressing a cause may be too subtle for Google to make the distinction between the two meanings.

We also found that Google was much more successful at translating S14 (scope of modifiers) than our NMT systems. This could be because it has a larger window of context and is better at processing the sentence as a whole, or because it has more training data that is likely to contain phrasings (i.e., examples of plausible combinations of these units) similar to the sentences that we created.

4.3.3.6. DeepL

DeepL was the strongest system overall, in both lexical and syntactic challenges. As summarized in Table 4, it managed to accurately translate 54.8% of the lexical challenge sentences and 65.0% of the syntactic challenge sentences. For some of the sentences that DeepL failed to translate acceptably (e.g., S4e and S4f), we found that—while they had to be marked as incorrect because we were only considering the challenging element—the sentences themselves require much less post-editing than the incorrect translations that other systems, such as our hybrid SMT system, generated for those same sentences.

DeepL's results for S14 (scope of modifiers) were also comparable to Google's, perhaps because the system also has a larger window of context and is also better at processing the sentence as a whole than our in-house systems. As well, the system was slightly stronger than other systems at translating homographs

All in all, we found that DeepL's translations of our sentences tended to be more representative of a native French speaker's translation, despite it failing some of our challenges.

4.4. Additional findings

Though the point of this challenge set was to test how different systems would perform when faced with specific lexical and syntactic difficulties, the analysis of our results also allowed us to identify recurring translation patterns found elsewhere in the sentences. Some of these patterns seem to be specific to a certain architecture, while others are found scattered through all of our systems. In this section, we will go over these findings and will try to interpret the systems' choices of translation.

4.4.1. Difficulties affecting multiple systems

Ambiguity, whether lexical or structural, was a recurrent difficulty that we noticed among multiple systems. For example, in S5a, we noticed that all of our NMT systems mistranslated *company* as *entreprise*, while we meant *company* as *companionship* and were expecting a solution such as *compagnie*. Similarly, in S5b, our NMT systems were not able to translate *make a living freelancing* with the missing *by* (*make a living by freelancing*). Instead, they all translated *living freelancing* as a noun phrase and, as a result, generated *un freelancing vivant*. In cases where systems rely on the meaning of a word in order to build a context around it, mistranslating a single word could lead to a system mistranslating other parts of the sentence.

This is not always the case, however, and it does not seem to apply to all systems. For instance, in S10b, all of our systems failed to identify *asked* as a past participle and translated it as the past tense verb. We initially thought that this could have contributed to our hybrid SMT system and our CNN-based system failing to translate the challenging element, but our RNN-based and attention-based systems successfully translated the challenging element regardless of their misinterpretation of *asked*.

While some of these cases of mistranslation could be due to ambiguity, in other cases, such as S8d and S8h, they could also be because the word (*wrinkly*) might not have been found in the training data, as our systems all struggled to translate the word. Our hybrid SMT system left the word in English in the target sentence, while our CNN-based system and our RNN-based system translated it as *ridicules*, and our attention-based system translated it as *troublés*. Although this word was inaccurately translated by all of

our systems, this sentence highlights the ability of NMT systems to potentially "guess" the translation based on other lexical linkages they may make in their black box, from the training data provided. SMT systems, on the other hand, will oftentimes simply retranscribe the unknown word.

Beyond lexical and syntactic elements, we also noticed that our use of an em dash in S10e could have thrown some of the systems off. While em dashes are commonly used in English to delimit a parenthetical element, commas will be used in French instead. Our CNN-based system and our attention-based system accurately translated the punctuation mark, whereas our hybrid SMT system and our RNN-based system, as well as Google and DeepL, used a hyphen in their translation.

These are all cases that are worth digging deeper into. However, they are beyond the scope of our challenge set analysis and would need to be explored in future work.

4.4.2. Difficulties affecting certain systems in particular

Through our analysis, we also found some recurring patterns specific to certain systems. These patterns, although unrelated to our challenge set per se, provide us with an insight into the systems' handling of certain element.

For example, our hybrid SMT system seemed to have trouble processing verbal contractions (S6e: *doesn't > doesn't*) and with apostrophes in general (S8g: *my aunt's old car > ma tante's vieille voiture*; S8h: *Katie's new house > Katie's nouvelle maison*), and as such, leaves them as is in the target sentence. We also noticed a number of agreement issues in our hybrid SMT system's translations; for example, in S5b, though it used the correct solution *bien que*, it did not use the required subjunctive mood after the conjunctive phrase. Another agreement issue example is in S7b, where it produced the sentence *des dizaines de tanières de loutres ont été complètement détruits*.

Our CNN-based system seems to often split sentences or add random capitalization mid-sentence. In S4e, for example, we noticed that the system tried to split the sentence into two by adding a capital mid-sentence (I noticed a *letter I* had never opened > j'ai remarqué une *lettre Je* n'avais jamais ouvert). We believe the system really is splitting the sentence and not just translating the capital *I* as a capitalized *Je*, since there have been other occurrences (such as in S3f: *nous sommes Les compiler*; and S4h:

dix-huit *Lettre* *d'une page*) where our CNN-based system made that error. We think this might be because we omitted the conjunction *that* (a letter *that* I had never opened), that would have been used to introduce the subordinate clause "I had never opened", which could have confused the system and would also explain the error in agreement in "Je n'avais jamais *ouvert*". The system may have interpreted the sequence as two separate sentences without a separating period and accordingly started a new sentence (although it did neglect to add a period to complete this adaptation). Based on this observation, we hypothesize that there could be a need for CNN-based systems to have explicit linkages between clauses in the source, especially since this error was not found in the other systems.

Additionally, we noticed that our CNN-based system once again capitalized certain words in S4h, as though to split the sentence, but in a seemingly less logical manner this time. The words *Le* (*souper*) and *Lettre* were capitalized and though they are both coincidentally words found before a punctuation mark in the source sentence, they are not the only words found before a punctuation mark (there was also *glass* > *verre*), which makes it impossible for us to draw any conclusions.

As for our RNN-based system, we found it sometimes omits part of the source sentence, as seen in S7f and S16g.

We also noticed that our attention-based system seems to produce more repetitions (it repeated *de la gare* four times in S5d and *très bien situé* twice in S5h). This is typical of NMT models and is an issue that researchers have raised as being a new problem that statistical models did not have (Park *et al*., 2020). After a quick search online, we found that other contributors to the OpenNMT Forum have experienced this problem with Transformer specifically. This issue can potentially be solved through "n-gram blocking" during the decoding stage. It essentially consists of discarding candidate translations that contain a repetitive sequence of tokens (Kulikov *et al*., 2019) to avoid awkward repetitions in the target sentence.

Our attention-based system was also the only system to create new words (*effondue* instead of *effondrée* in S7g, *cristé* in S14h). In theory, all of our NMT systems should have been able to create new words, as they all went through the same tokenization process, allowing them room for an open vocabulary. Therefore, we were

surprised to see that our attention-based system was the only one to "make use" of that capacity.

4.5. Recommendations

Based on the detailed analysis of our results provided in Sections 4.1 and 4.2, and on the findings we identified in 4.3 and 4.4, we were able to establish a (non-exhaustive) list of recommendations. In 4.5.1, we will be presenting possible changes that can be made to our challenge set in order to reduce ambiguity and allow us to focus on the challenging element. Then, in 4.5.2, we will put forward some recommendations to guide users of MT towards the system(s) that might best fit their needs, as well as provide developers of MT systems with potential tweaks to improve their systems's performance according to the results obtained through our challenge set.

4.5.1. Possible improvements to the challenge set

While we did our best to create sentences that we thought were unambiguous, some of the systems' solutions allowed us to see elements we may have overlooked, but that were proven to be problematic for certain systems to translate. In this section, we discuss modification of certain sentence structures or wordings in our challenge set.[46]

As mentioned in the previous section, the fact that we omitted the relative pronoun in S4e may have made it more difficult for our CNN-based system to process the sentence. As such, we believe it would be best to have explicit linkages between clauses in all of our source sentences.

The CNN system also had recurrent problems with contractions, often leaving them untranslated or partially translated; we also hypothesized that the system may have missed a cue for disambiguation in a verb tense in S3f because of its inability to properly interpret a contraction. In future work, the contractions in the challenge set sentences could be replaced by full forms to avoid these problems.

Both of the above measures are recommended in some guidelines for writing for (machine) translation. While not necessarily reflective of "real world" examples, this

[46] A copy of both the original challenge set and the revised version will be made available on https://github.com/CoralineDoan/challenge-set

would help to avoid errors outside the challenge item that might contribute to problems affecting the ability to evaluate systems' handling of the item in question. Such issues could then be targeted for evaluation in separate challenges.

We also noticed that some words that we did not think would be considered "rare" actually caused unexpected translations. These words may not have been found in the training data, resulting in odd translations (e.g., *wrinkly* > *ridicules* [CNN-based] or *troublés* [attention-based]). A more thorough review of the systems' output to identify these odd translations and to replace their corresponding source with a more common word might facilitate the translation of the challenge itself, or at least reduce the amount of post-editing required on the sentence as a whole. With direct access to the training data, we might also have been able to adopt a strategy similar to that used by Isabelle, Cherry, & Foster (2017, p. 2487), setting a minimum threshold of occurrences in the training data to select our vocabulary for the challenge set.

Furthermore, in trying to include different types of difficulties in the challenges and maintain the required characteristics for the challenge set sentences, we sometimes accidentally ignored other elements that added complexity (e.g., ambiguity) to the sentence. For example, in S16c, while trying to reach the minimum number of words established for our short sentences, we unintentionally added more than one noun as interruption in the sentence (*He put his brother's photo in a book after breaking the frame it was in*). A similar case was noted in S14c. In S1g, we included more than one temporal conjunction, which appeared to have caused problems for some systems. In a few cases, challenges such as part-of-speech or semantic ambiguity in the sentences may have had unintended consequences for the systems' interpretation of the challenges (e.g., *cooks* in S3d, *dated* in S9g). Going forward, we would ensure that our number of interruptions and other potentially challenging phenomena is more consistent (when these are required) and/or eliminated (if they are not intended) to ensure a more coherent evaluation.

Depending on needs, and in light of these initial observations and the hypotheses we have made here, we might also design sentences to more precisely measure the impact of placement or relative proximity of different types of items (e.g., antecedents and anaphoric references, repetitions resulting from asymmetrical equivalence), or the nature

of cues for disambiguation in the sentences (e.g., in antecedents or in other sentence elements), to more fully appreciate the subtleties of possible influences. For example, given our observation that the different placement of some of our ambiguous items may affect performance for some systems, it might be interesting to test variants of the same sentence with the challenge set item placed in different locations (e.g., S6ai - *While Megan loves a good steak, her brother does not eat meat* and S6aii - *Megan loves a good steak, while her brother does not eat meat*). Similarly, there may be potential to target the effect of sense frequency on the systems' choice of equivalent for an ambiguous lexical item (e.g., by including challenges with for very frequent, moderately frequent, and rare senses of certain lexical items), to better understand how frequency interacts with other factors (e.g., contextual cues) in the systems' attempts to resolve the ambiguities.

Overall, despite the challenges for this project, observing these issues in the data has allowed us to identify some additional areas for exploration in a larger and more developed challenge set, which would provide a more comprehensive portrait of possible issues for NMT and an opportunity to do more developed statistical analyses if required.

4.5.2. Choice of systems and potential tweaks

Through our findings, we were able to provide some recommendations, both for users of MT and developers of MT systems. We found that the different systems in our experiment seem to be able to handle different types of challenges, as outlined in Table 39.

		First	Second	Third	Fourth	Fifth	Last
Lexical	Semantic ambiguity	*Google*	*DeepL*	Attention	RNN	Hybrid SMT CNN	
	Asymmetrical equivalence	*DeepL*	Hybrid SMT	RNN Attention	CNN *Google*		
Syntactic	Scope	*Google* *DeepL*	Hybrid SMT	RNN Attention	CNN		
	Anaphora	*DeepL*	*Google*	RNN	Hybrid SMT	CNN	Attention

Table 39 – Ranking of the systems by challenge type

We found that our RNN-based system often fell in the third and fourth positions in every challenge type, making it a potential "average" architecture, neither excelling nor failing, while other systems seem to have their strengths and weaknesses (with the exception of our CNN-based system, which seems to handle most challenges less well than the other systems, as we expected).

With the ranking provided, we are hoping users of MT can tailor their system choice, as well as their development of writing and postediting guidelines to handle the challenges they think they would be likely to encounter.

When it comes to postediting outputs, we recommend that posteditors be attentive to the issues identified through this challenge set, similarly to how translators would be attentive to the translation difficulties identified in literature. However, the handling of the various challenges in this study may suggest some areas for particular attention in some systems. For example, users who opt for an RNN-based system may need to be attentive to all the challenging elements outlined in our challenge set, but may spend less time postediting them, as errors may not be as frequent as in some other systems. Conversely, users who decide to use an attention-based system might need to pay particular attention to anaphora, but less to the other difficulties identified in this challenge set. As for users who choose a hybrid SMT system, they might be able to target longer-range dependencies for particular attention and worry less about closer ones. A hybrid SMT system may also be a good choice if the texts that require translation involve collocations or local word associations and co-occurrences, such as *consumed with guilt*, *burning with hatred*.

In terms of writing for translation, we recommend avoiding having multiple temporal conjunctions present in the sentence if one is using an attention-based system that needs to tackle ambiguous lexical items. We also recommend keeping the sentences at a certain number of words minimum when using an RNN-based system, as it seems to work better when provided with a larger context of window. Interestingly, this recommendation (based on the performance of our systems on long and short sentences) somewhat contradicts the conventional guidelines for writing for translation, which typically recommend keeping sentences short. Users may wish to monitor their chosen systems' performance closely to determine optimal sentence length for their purposes.

Furthermore, by looking at the results obtained (illustrated in Figure 7, which presents the contribution of each challenge type to the number of correct translations obtained by each system), developers can think about whether or not tweaks could be useful and where they would be needed. Improving how our attention-based system translated *these* could be one focus, perhaps by adding more training data where the proximity aspect is properly expressed.

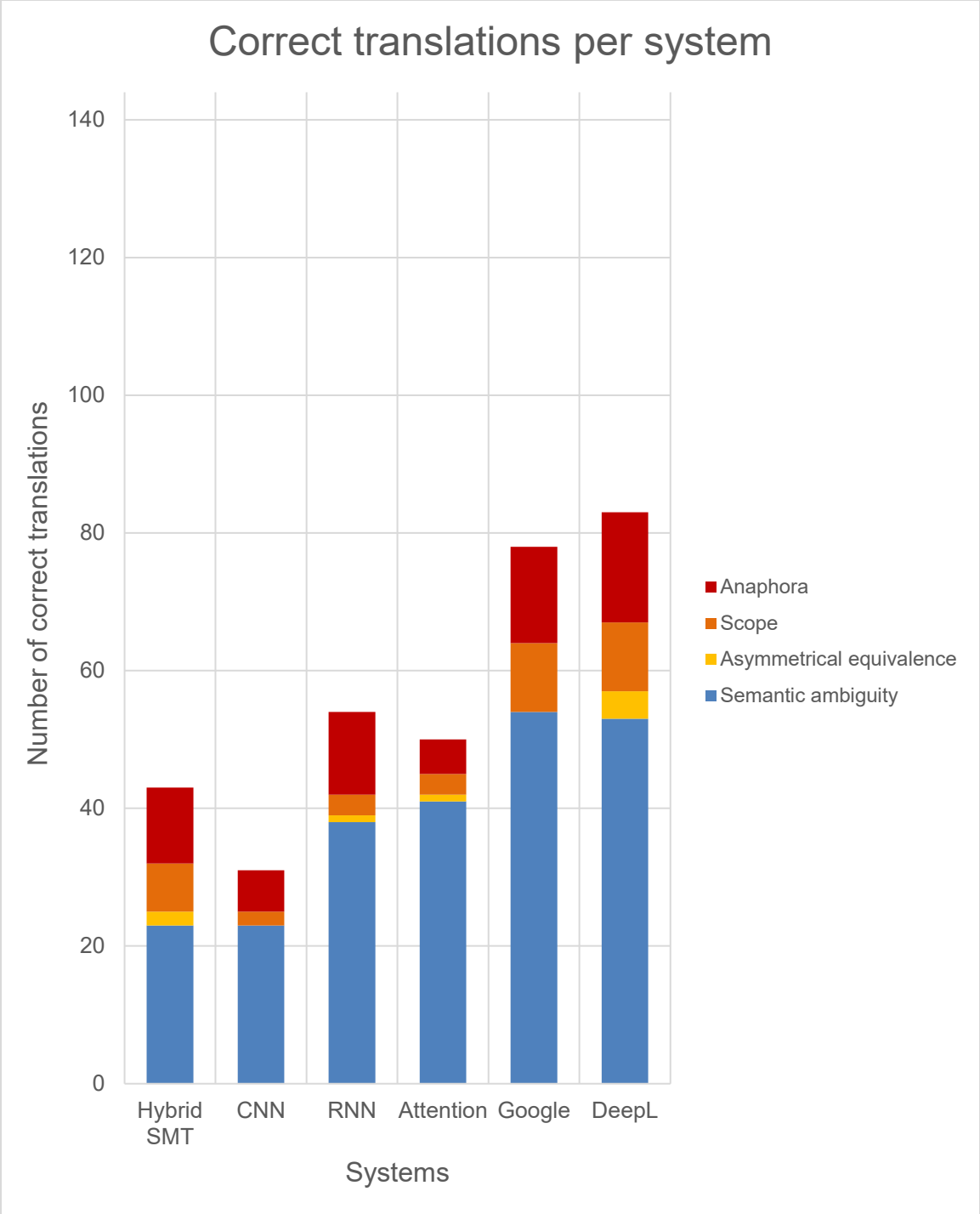

Figure 7 – Number of correct translations per system

Conclusion

The goal of this project was to get a sense of how different encoder-decoder NMT models will perform, based on their architecture, when faced with different challenges, through the lens of a translator. Based on our background and on our work experience in translation, we developed a challenge set for English to French translation (which can be reused), adopting a bottom-up approach, where lexical and syntactic difficulties were isolated. We assessed all the sentences produced by our systems and we were able to identify some patterns. With the different categories included in our challenge set and with the different variants that we developed, we were in a position to say which model performed better in which context.

We had two hypotheses: first, that different NMT architectures would perform differently when faced with different types of challenges, and, second, that a CNN-based system would most likely achieve poor results in a challenge set, since it would tend to struggle with accuracy in long-range dependencies and challenge set evaluations rely on accuracy (as opposed to fluency). Through our results and analysis, we did identify different patterns that seem specific to each of the architectures and, as hypothesized, our CNN-based system had the lowest proportion of sentences correctly translated.

For example, when faced with ambiguous lexical items, we found that our CNN-based system tends to sometimes ignore the challenging element altogether, though we cannot explain, based on its architecture, why it did so. We also found that it sometimes produced what seemed to be a partial solution (e.g., *que* instead of *alors que*). This corresponds to what we know of CNN-based systems, as they work by focusing on a particular section of the source sentence at a time and do not consider the sentence as a whole (see Section 1.1.5.1.2 for more details). As such, they might be more inclined to produce translations that seem "choppier". Results for our RNN-based system were more surprising, as we expected this system to do better with an ambiguous lexical item placed mid-sentence, rather than at the beginning of the sentence. Since RNN-based systems process sentences in a sequential manner (see 1.1.5.1.1), we thought that having the linking word at the beginning of the sentence might be harder to translate because the system does not have the required information or cue to establish the most likely

meaning. Yet, it was not always the case and we found that the results varied according to the meaning rather than the position of the ambiguous lexical item. For our attention-based system, we found that it mostly struggled to translate ambiguous lexical items when there were multiple temporal conjunctions present in the sentence.

When we tested for asymmetrical equivalence, we noticed that our hybrid SMT system had better results than all of our NMT systems, though this could be because systems are unable to detect what we humans consider awkward wordings.

We observed more homogenous results in our syntactic difficulties. In terms of scope, our three NMT systems performed similarly, with our RNN-based and attention-based systems obtaining slightly higher results than our CNN-based system, as expected and as established in one of our hypotheses. With anaphora, however, our CNN-based system unexpectedly had better results than our attention-based system, which failed to accurately translate all of our sentences testing *these*. Nevertheless, our RNN-based system was the system to perform the best among our experimental systems, translating 50.0% of the sentences testing for anaphora accurately.

While we made the decision to test our systems using a challenge set approach, we recognize that there are limitations to this method. Challenge sets are useful when we want to look at a specific phenomenon and highlight a system's particular strength or weakness, but they do not provide insight in a system's overall performance and translation quality. We realize that, in some of our translations, although the challenge was marked as having been successfully tackled, the rest of the sentence could still be flawed. To get an accurate representation of how a model performs when faced with a certain difficulty, we would need to combine this challenge set approach with another type of machine translation evaluation, such as a natural test set, as suggested by Popović and Castilho (2019).

Furthermore, since challenge sets need to be developed and evaluated by humans, it is not easily possible to have a large enough number of sentences for results to be statistically significant.

In creating this challenge set and analyzing its results, we were limited in terms of available resources. The only evaluators being ourselves, we had to limit the number of sentences we created and tested, and some of our results might not be representative of

the systems' overall performance, as we only have a limited sample of test sentences. It would be interesting to further develop, refine and/or complement our challenge set.

In some categories of our challenge set, we observe great variability between the individual examples, which made it difficult for us to identify any patterns. Human errors and oversight may also occur: in retrospect, we realized that some of our systems' failure to accurately translate the challenging element might have been caused by another element, found elsewhere in the sentence. This was not intentional and some of our sentences could be improved in future work.

We realize we were limited by time (and space) for this research project and, therefore, have more we hope to achieve in the future to improve our challenge set and to dive deeper into our analysis of the translations we have obtained. Potential work that could be pursued in the future includes a more fine-grained study of the effect of the positioning of certain ambiguous lexical items. By using variants of a single sentence with different positioning of the item to be tested, we could test our hypothesis that perhaps some senses might be more likely to occur in certain positions in a sentence and are thus, more likely to be correctly recognized by the systems. This research could help inform improvement of MT systems through added rules, or help inform writing for translation.

Based on the preliminary observations of our lexical ambiguities, there also appears to be potential in exploring possible correlation between the variations in frequency of a sense and how various systems translate it.

It would also be worth looking into the sentences that we found hard to classify as being correct or incorrect. In cases where it was difficult to make a clear decision (cases marked by an asterisk in our results), we might want to establish a more uniform, extensive, and overall easier evaluation process, and have more human evaluators deliberate on the ratings. Although we did our best to establish a rigorous evaluation method before putting our sentences to the test, some systems still generated solutions that we did not anticipate, and therefore could not have taken into account when developing our evaluation method. (For example, should cases where the challenging element was ignored but the sentence produced still makes sense be categorized as correct? Should there be more than the two correct and incorrect categories?)

These are all considerations we were able to establish while conducting our work and we hope to either one day address them or see them addressed. This said, adopting a challenge set approach is still relevant for MT evaluation because—although it does not provide us with the whole picture and will not reflect the time and effort required by translators to postedit sentences produced by NMT systems— it will reflect elements that translators need to look out for should they opt to use models based on these architectures.

Works cited

Agrawal, R., Turchi, M., & Negri, M. (2018). Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. 11-20.

Arnold, D. J. (2003). Why translation is difficult for computers. In Harold Somers (Ed.), *Computers and Translation: A translator's guide* (pp. 119-142). John Benjamins.

Banchs, R. E. (2016, November 1). *Continuous Vector Spaces for Cross-Language NLP Applications*. [Tutorial]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Austin, Texas. https://www.aclweb.org/mirror/emnlp2016/tutorials/banchs-t5.pdf

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural *versus* Phrase-Based Machine Translation Quality: a Case Study. Retrieved from https://aclweb.org/anthology/D16-1025

Bergen, M. & Wagner, K. (2015). Welcome to the AI Conspiracy: The 'Canadian Mafia' Behind Tech's Latest Craze. Retrieved from https://www.recode.net/2015/7/15/11614684/ai-conspiracy-the-scientists-behind-deep-learning

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J., Jelinek, F., Lafferty, J. D., … Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79-85. Retrieved from https://aclanthology.info/pdf/J/J90/J90-2002.pdf

Brown, P. F., Pietra, S. D., Pietra, V. J., & Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), 263-311. Retrieved from http://www.aclweb.org/anthology/J93-2003

Bowker, L. & Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies, 10*(2), 165-186. https://doi.org/10.1075/tis.10.2.01bow

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249-256.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics, 108*, 109-120. https://doi.org/10.1515/pralin-2017-0013

Castilho, S., Popović, M., & Way, A. (2020). On Context Span Needed for Machine Translation Evaluation. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3735-3742.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8). Retrieved from https://arxiv.org/pdf/1409.1259.pdf

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014b). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Retrieved from https://arxiv.org/pdf/1406.1078.pdf

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Retrieved from https://arxiv.org/pdf/1412.3555.pdf

Delisle, J. (1993). La traduction raisonnée – Livre du maître : méthode par objectifs d'apprentissage. Les Presses de l'Université d'Ottawa.

Delisle, J. & Fiola, M. A. (2013). *La traduction raisonnée : Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. Les Presses de l'Université d'Ottawa.

Delisle, J. & René, A. (2003). *La traduction raisonnée : Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. Les Presses de l'Université d'Ottawa.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1370-1380, http://dx.doi.org/10.3115/v1/P14-1129

Doherty, S., O'Brien, S., & Carl, M. (2010). Eye Tracking as an Automatic MT Evaluation Technique. *Machine Translation*, *24*(1), 1-13. http://dx.doi.org/10.1007/s10590-010-9070-9

Domhan, T. (2018). How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 1799-1808. https://doi.org/10.18653/v1/P18-1167

Doshi, K. (2021, January 16). Transformers Explained Visually (Part 3): Multi-head Attention, deep dive. *Towards data science*. https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853

Dubuc, R. (2002). *Manuel pratique de terminologie 4e éd*. Linguatech Éditeur inc.

Ferguson, N. (2019, February 6). The Past and Present of Translation Memory Technology. *SDL Trados*. https://www.trados.com/blog/past-present-translation-memory-technology.html

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291-309. https://doi.org/10.1075/ts.6.2.06for

Forcada, M. L., Scarton, C., Specia, L., Haddow, B., & Birch, A. (2018). Exploring Gap Filling as a Cheaper Alternative to Reading Comprehension Questionnaires when Evaluating Machine Translation for Gisting. *Proceedings of the Third Conference on Machine Translation (WMT) (Volume 1: Research Papers)*, 192-203. http://dx.doi.org/10.18653/v1/W18-6320

García, I. (2006). Translators on translation memories: a blessing or a curse? In A. Pym, A. Perekrestenko & B. Starink (Eds.), Translation Technology and its Teaching (pp. 97-105). Intercultural Studies Group Universitat Rovira i Virgili.

García, I. (2009). Beyond Translation Memory: Computers and the Professional Translator. *The Journal of Specialised Translation*, *issue 12*. Retrieved from https://www.jostrans.org/issue12/art_garcia.pdf

Gaspari, F., Toral, A., Naskar, S. K., Groves, D., & Way, A. (2014). Perception vs. reality: measuring machine translation post-editing productivity. *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, 60-72. Retrieved from https://aclanthology.org/2014.amta-wptp.5.pdf

Geitgey, A. (2016). Machine Learning is Fun Part 5: Language Translation with Deep Learning and the Magic of Sequences. *Medium*. Retrieved from https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa

Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2017). A Convolutional Encoder Model for Neural Machine Translation. Retrieved from https://arxiv.org/pdf/1611.02344v3.pdf

Gotti, F., Langlais, P., & Lapalme, G. (2014). Designing a machine translation system for Canadian weather warnings: A case study. *Natural Language Engineering*, *20*(3). 399-433. https://doi.org/10.1017/S135132491300003X

Guillou, L. & Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 636-643. Retrieved from https://www.aclweb.org/anthology/L16-1100

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufmann Publishers.

Hardmeier, C. & Guillou, L. (2018). Pronoun Translation in English-French Machine Translation: An Analysis of Error Types. Retrieved from https://arxiv.org/pdf/1808.10196.pdf

He, W., He, Z., Wu, H., & Wang, H. (2016). Improved Neural Machine Translation with SMT Features. Retrieved from http://research.baidu.com/Public/uploads/5acc2bb7a7cf8.pdf

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., & Post, M. (2017). Sockeye: A Toolkit for Neural Machine Translation. Retrieved from https://arxiv.org/pdf/1712.05690v1.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. Retrieved from http://www.bioinf.jku.at/publications/older/2604.pdf

Hutchins, J. (1995). Machine Translation: A Brief History. In E. F. K. Koerner et R. E. Asher (Eds.), *Concise History of the Language Sciences: from the Sumerians to the Cognitivists* (431-445). Oxford: Pergamon Press. Retrieved from http://hutchinsweb.me.uk/ConcHistoryLangSci-1995.pdf

Hutchins, J. (2001). Machine translation over fifty years. *Histoire, Epistémologie, Langage*, *23*, 7-33. https://doi.org/10.3406/hel.2001.2815

Hutchins, J. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. In R. E. Frederking, K. B. Taylor (Eds.), *Machine Translation: From Real Users to Research, 6th AMTA conference* (102-114). Berlin: Springer, Berlin, Heidelberg. Retrieved from http://www.hutchinsweb.me.uk/AMTA-2004.pdf

Hutchins, J. (2005). Example-Based Machine Translation: A Review and Commentary. *Machine translation*, 19(3/4), 197–211.

Hutchins, J. & Somers, H. L. (1992). *An Introduction to Machine Translation*. Retrieved from http://www.hutchinsweb.me.uk/IntroMT-TOC.htm

Isabelle, P., Cherry, C., & Foster, G.F. (2017). A Challenge Set Approach to Evaluating Machine Translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2486-2496. Retrieved from https://arxiv.org/pdf/1704.07431.pdf

Isabelle, P., Goutte, C., & Simard, M. (2007). Domain adaptation of MT systems through automatic post-editing. *Proceedings of Machine Translation Summit XI*, 255-261. Retrieved from https://nrc-publications.canada.ca/eng/view/accepted/?id=c941bc66-9dbe-495d-af7e-039c82873e21

Isabelle, P., & Kuhn, R. (2018). A Challenge Set for French → English Machine Translation. Retrieved from https://arxiv.org/pdf/1806.02725.pdf

Joty, S., Durrani, N., Sajjad, H., & Abdelali, A. (2017). Domain adaptation using neural network joint model. *Computer Speech & Language*, *45*, 161-179. https://doi.org/10.1016/j.csl.2016.12.006

Kay, M. (2003). The Proper Place of Men and Machines in Language Translation. In S. Nirenburg, H. L. Somers & Y. A. Wilks (Eds.), Readings in Machine Translation (pp. 221-232). The MIT Press. https://doi-org.proxy.bib.uottawa.ca/10.7551/mitpress/5779.001.0001

Kenny, D. (2018). Machine Translation. In P. Rawling & P. Wilson (Eds.), The Routledge handbook of translation and philosophy (pp. 428-445). Routledge.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Retrieved from https://arxiv.org/pdf/1408.5882.pdf

Kit, C. & Wong, B. T.-M. (2014). Evaluation in machine translation and computer-aided translation. In Sin-Wai Chan (Ed.), *Routledge Encyclopedia of Translation Technology* (pp. 213-236). Routledge.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit*, 79-86

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, *23*, 241-263. https://doi.org/10.1007/s10590-010-9076-3

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, P. & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28-39. http://dx.doi.org/10.18653/v1/W17-3204

Koehn, P. & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. *Proceedings of the Workshop on Statistical Machine Translation*, 102-121.

Koponen, M. & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *The Journal of Specialised Translation*, (23), 118-136.

Kulikov, I., Miller, A. H., Cho, K., & Weston, J. (2019). Importance of Search and Evaluation Strategies in Neural Dialogue Modeling. Retrieved from https://arxiv.org/pdf/1811.00907.pdf

L'Homme, M.-C. (2008). *Initiation à la traductique*. Linguatech Éditeur inc.

Lakew, S.M., Cettolo, M., & Federico, M. (2018). A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. *Proceedings of the 27th International Conference on Computational Linguistics*, 641-652.

Lardinois, F. (2010, September 9). Linguee Brings Translation Dictionaries into the 21st Century. Readwrite. https://readwrite.com/2010/09/09/linguee_online_translation_dictionary_english_spanish_german_french/

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98-113. https://doi.org/10.1109/72.554195

Le, Q. V., & Schuster, M. (2016, September 27). A Neural Network for Machine Translation, at Production Scale. Google AI Blog. https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

LeBlanc, M. (2013). Translators on translation memory (TM). Results of an ethnographic study in three translation services and agencies. *Translation & Interpreting*, *5*(2), 1-13. http://dx.doi.org/10.12807/ti.105202.2013.a01

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Longman Dictionary of Contemporary English. (n.d.). As. In *Longman Dictionary of Contemporary English Online*. Retrieved May 24, 2020, from https://www.ldoceonline.com/dictionary/as

Longman Dictionary of Contemporary English. (n.d.). When. In *Longman Dictionary of Contemporary English Online*. Retrieved May 24, 2020, from https://www.ldoceonline.com/dictionary/when

Longman Dictionary of Contemporary English. (n.d.). While. In *Longman Dictionary of Contemporary English Online*. Retrieved May 24, 2020, from https://www.ldoceonline.com/dictionary/while

Longman Dictionary of Contemporary English. (n.d.). With. In *Longman Dictionary of Contemporary English Online*. Retrieved May 24, 2020, from https://www.ldoceonline.com/dictionary/with

Luong, M.-T., Pham, H. & Manning, C. D. (2015a). Effective approaches to attention-based neural machine translation. Retrieved from https://arxiv.org/pdf/1508.04025.pdf

Luong, M.-T., Sutskever, I., Le, Q., Vinyals, O., & Zaremba, W. (2015b). Addressing the rare word problem in neural machine translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 11-19. http://dx.doi.org/10.3115/v1/P15-1002

Macken, L., Prou, D., & Tezcan, A. (2020). Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process. *Informatics 2020*, *7*(2), https://doi.org/10.3390/informatics7020012

Marshman, E. (2014). Taking Control: Language Professionals and Their Perception of Control when Using Language Technologies. *Meta*, *59*(2), 380-405. https://doi.org/10.7202/1027481ar

Matusov, E. (2019). The Challenges of Using Neural Machine Translation for Literature. *Proceedings of the Qualities of Literary Machine Translation*, 10-19. Retrieved from https://aclanthology.org/W19-7302.pdf

Merriam-Webster. (n.d.). 'Since' vs. 'As' vs. 'Because'. *Merriam-Webster*. https://www.merriam-webster.com/words-at-play/since-as-because-usage

Merriam-Webster. (n.d.). As. In *Merriam-Webster.com dictionary*. Retrieved May 8, 2020, from https://www.merriam-webster.com/dictionary/as

Merriam-Webster. (n.d.). When. In *Merriam-Webster.com dictionary*. Retrieved May 8, 2020, from https://www.merriam-webster.com/dictionary/when

Merriam-Webster. (n.d.). While. In *Merriam-Webster.com dictionary*. Retrieved May 8, 2020, from https://www.merriam-webster.com/dictionary/while

Merriam-Webster. (n.d.). With. In *Merriam-Webster.com dictionary*. Retrieved May 8, 2020, from https://www.merriam-webster.com/dictionary/with

Novikova, J., Dušek, O., Curry, A. C., Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2241–2252. http://dx.doi.org/10.18653/v1/D17-1238

Papineni, K., Roukos, S., Ward, T., & Zhu, W-J. (2002). *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. Retrieved from https://www.aclweb.org/anthology/P02-1040.pdf

Park, C., Yang, Y., Park, K., Lim, H. (2020). Decoding Strategies for Improving Low-Resource Machine Translation. *Electronics*, *9(10)*, 1562. https://doi.org/10.3390/electronics9101562

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: J. Moorkens, S. Castilho, F. Gaspari, S. Doherty (Eds.), *Translation Quality Assessment* (pp. 129-158). Springer, Cham. https://doi.org/10.1007/978-3-319-91241-7_7

Popović, M. (2021). On nature and causes of observed MT errors. *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: MT Research Track)*, 163-175. Retrieved from https://aclanthology.org/2021.mtsummit-research.14.pdf

Popović, M., & Castilho, S. (2019). Challenge Test Sets for MT Evaluation. *Proceedings of Machine Translation Summit XVII (Volume 3: Tutorial Abstracts)*. Retrieved from https://www.aclweb.org/anthology/W19-7602

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation (WMT), (Volume 1: Research Papers)*, 186-191. https://doi.org/10.18653/v1/W18-6319

Pouget-Abadie, J., Bahdanau, D., van Merrienboer, B., Cho, K., & Bengio, Y. (2014). Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 78-85. http://dx.doi.org/10.3115/v1/W14-4009

Public Works and Government Services Canada. (2017, August 16). *Translation Bureau – Language Comprehension Tool*. Public Services and Procurement Canada's GCintranet. https://outilta-mttool.spac-pspc.gc.ca/index-eng.php

Quah, C. K. (2006). Computer-Aided Translation Tools and Resources. *Translation and technology* (4). Retrieved from https://ebookcentral.proquest.com

Reiter, E. (2018). A Structured Review of the Validity of BLEU. *Computational Linguistics*, *44*(3), 393–401. http://dx.doi.org/10.1162/coli_a_00322

Renaud-Bray. (n.d.). La Traduction raisonnée 3e éd. Renaud-Bray. https://www.renaud-bray.com/Livres_Produit.aspx?id=1448541&def=Traduction+raisonn%c3%a9e(La)+3e+%c3%a9d.%2cDELISLE%2c+JEAN%2c9782760308060

Rikters, M. (2018). Impact of Corpora Quality on Neural Machine Translation. Retrieved from https://arxiv.org/pdf/1810.08392.pdf

Roturier, J. (2006). *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users*. [Doctoral dissertation, Dublin City University] Retrieved from http://doras.dcu.ie/18190/1/Johann_Roturier_20130116094552.pdf

Sénécal, A. (2014). La théorie en prise directe sur la pratique. *Circuit, le magazine d'information des langagiers*, *122*. https://www.circuitmagazine.org/la-theorie-en-prise-directe-sur-la-pratique

Sennrich, R., Haddow, B., Birch, A. (2016) Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725, http://dx.doi.org/10.18653/v1/P16-1162

Snover, M., Madnani, N., Dorr, B. J., & Schwartz, R. (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 259-268.

Sumita, E., & Iida, H. (1992). Example-Based NLP Techniques – A Case Study of Machine Translation. *Proceedings of Statistically-Based NLP Techniques Workshop (AAAI'92)*, 81-88.

Sun, S. (2019). Measuring Difficulty in Translation and Post-editing: A Review. In D. Li, V. Lei, & Y. He (Eds.), *Researching Cognitive Processes of Translation* (pp. 139-168). Springer. https://doi.org/10.1007/978-981-13-1984-6

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Retrieved from https://arxiv.org/pdf/1409.3215.pdf

Systran. (n.d.) *What is Machine Translation? Rule Based Machine Translation vs. Statistical Machine Translation*. Systran beyond language. https://www.systransoft.com/systran/translation-technology/what-is-machine-translation/

Tang, G., Müller, M., Rios, A., Sennrich, R. (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4263-4272.

Tu, Z., Liu, Y., Lu, Z., Liu, X., & Li, H. (2017). Context Gates for Neural Machine Translation. Retrived from https://arxiv.org/pdf/1608.06043.pdf

Torregrosa, D., Pasricha, N., Chakravarthi, B. R., Masoud, M., Arcan., M., Alonso, J., & Casas, N. (2019). Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models. *Proceedings of MT Summit XVII (Volume 2)*, 125-133. Retrieved from https://www.aclweb.org/anthology/W19-6725.pdf

Uszkoreit, J. (2017, August 31). Transformer: A Novel Neural Network Architecture for Language Understanding. Google AI Blog. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

V V, P. (2016). Bayesian Regularization for #NeuralNetworks. *Medium*. Retrieved from https://medium.com/autonomous-agents/bayesian-regularization-for-neuralnetworks-2f2d34f03adc

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998-6008.

Wang, C., & Sennrich, R. (2020). On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. Retrieved from https://arxiv.org/pdf/2005.03642.pdf

Wang, X., Lu, Z., Tu, Z., Li, H., Xiong, D., & Zhang, M. (2016). Neural Machine Translation Advised by Statistical Machine Translation. Retrieved from https://arxiv.org/pdf/1610.05150.pdf

White, J. S., O'Connell, T. & O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 193-205.

Wu, Y., Schuster, M., Chen, Z., Le, Q. L., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Retrieved from https://arxiv.org/pdf/1609.08144.pdf

Yang, Z., Cheng, Y., Liu, Y., & Sun, M. (2019). Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6191-6196. http://dx.doi.org/10.18653/v1/P19-1623

Yarats, D., Gehring, J. & Auli, M. (2017). A novel approach to neural machine translation. Retrieved from https://engineering.fb.com/2017/05/09/ml-applications/a-novel-approach-to-neural-machine-translation/

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. Retrieved from https://arxiv.org/pdf/1702.01923.pdf

Ziganshina, L. E., Yudina, E.V., Gabdrakhmanov, A.I., & Ried, J. (2021). Assessing Human Post-Editing Efforts to Compare the Performance of Three Machine Translation Engines for English to Russian Translation of Cochrane Plain Language Health Information: Results of a Randomised Comparison. *Informatics 2021*, *8*(9). https://doi.org/10.3390/informatics8010009

Appendix A: Challenge set – Lexical difficulties

Ambiguity in the source language

"As" expressing simultaneity

Possible solutions: *alors que*, *quand*, or *au moment où*

Short sentences		
"as" before the propositions		
S1a	Source	**As** I took my fries out of the bag, a seagull tried to steal some.
	Ref	**Au moment où** j'ai sorti mes frites du sac, une mouette a essayé de m'en voler.
✗	Hybrid SMT	**Comme** je l' ai pris mon fries du sac, un goéland tenté de voler certains.
✓	CNN	**Alors que** j' ai pris mes alevins du sac, un mouton a essayé de voler certains.
✗	RNN	**Comme** j' ai pris mes frites hors du sac, un mouette a essayé de voler certains.
✓	Attention	**Lorsque** j' ai retiré mes frites du sac, un alpin a tenté de les voler.
✓	Google	**Alors que** je sortais mes frites du sac, une mouette a essayé d'en voler.
✓	DeepL	**Alors que** je sortais mes frites du sac, une mouette a essayé d'en voler.
S1b	Source	**As** I put my curtains up, I realized the rod was crooked.
	Ref	**Quand** j'ai installé mes rideaux, je me suis rendu compte que la tringle était croche.
✗	Hybrid SMT	**Comme** j' ai mis mes rideaux en place, je me suis rendu compte que la tige était tordu.
✗*	CNN	J' ai mis mes rideaux, j' ai réalisé que la tige était tombée.
✓	RNN	**Lorsque** j ' ai mis mes rideaux, j ' ai constaté que la tige était collée.
✓	Attention	**Au moment où** j' ai mis mes rideaux, j' ai réalisé que la tige était étouffée.
✓	Google	**En remontant** mes rideaux, j'ai réalisé que la tringle était tordue.
✓	DeepL	**Alors que** je montais mes rideaux, je me suis rendu compte que la tringle était tordue.

"as" between the propositions		
S1c	Source	My knees cracked **as** I stood up.
	Ref	Mes genoux ont craqué **quand** je me suis levé.
✗	Hybrid SMT	Mes genoux fêlés **comme** je me suis levé.
✗	CNN	Mes genoux se sont fissurés **comme** je l' étais.
✓	RNN	Mes genoux se fissuraient **lorsque** j' étais monté.
✗	Attention	Mes genoux se sont fissurés **au fur et à mesure que** j' étais allé.
✓	Google	Mes genoux craquèrent **alors que** je me levais.
✓	DeepL	Mes genoux ont craqué **quand** je me suis levé.
S1d	Source	I discovered a hidden image **as** I was restoring an old painting.
	Ref	J'ai découvert une image cachée **alors que** je restaurais une ancienne peinture.
✗	Hybrid SMT	J' ai découvert une image caché que j' étais à la restauration d' une ancienne peinture.
✗	CNN	J' ai découvert une image cachée que j' étais en train de restaurer une vieille peinture.
✓	RNN	J' ai découvert une image cachée **alors que** je rétablissais une vieille peinture.
✓	Attention	J' ai découvert une image cachée **alors que** j' étais en train de restaurer une vieille peinture.
✓	Google	J'ai découvert une image cachée **alors que** je restaurais un vieux tableau.
✓	DeepL	J'ai découvert une image cachée **alors que** je restaurais un vieux tableau.
Long sentences		
"as" before the propositions		
S1e	Source	**As** I took my fries and ketchup out of the paper bag today at the beach, a vicious seagull tried to steal some of the fries.
	Ref	**Alors que** je sortais mes frites et mon ketchup du sac de papier aujourd'hui à la plage, une mouette vicieuse a essayé de voler quelques frites.
✗	Hybrid SMT	**Comme** j' ai pris mon fries et ketchup hors du sac de papier aujourd' hui à la plage, un goéland vicieuse tenté de voler certains des frites.

✗	CNN	**Comme** j' ai pris mes frites et ketchup à l' extérieur du sac en papier aujourd' hui à la plage, un marin vicieux a essayé de voler certains des frites.
✓	RNN	**Alors que** j' ai pris mes frites et que j' ai sorti du sac de papier aujourd' hui à la plage, un marin vicieux a essayé de voler certaines frites.
✓	Attention	**Alors que** j' ai retiré mes frites et mes ketchup du sac à papier aujourd' hui à la plage, un vicieux alpin a essayé de voler certains des frites.
✓	Google	**Alors que** je sortais mes frites et mon ketchup du sac en papier aujourd'hui à la plage, une mouette vicieuse a essayé de voler des frites.
✓	DeepL	**Alors que** je sortais mes frites et mon ketchup du sac en papier aujourd'hui à la plage, une mouette vicieuse a essayé de voler quelques unes des frites.
S1f	Source	**As** I put my blue and white curtains up, I realized the matte black metal rod was crooked and didn't match with the rest of the room.
	Ref	**Quand** j'ai installé mes rideaux bleu et blanc, je me suis rendu compte que la tringle en métal noir mat était croche et n'allait pas avec le reste de la pièce.
✗	Hybrid SMT	**Comme** j' ai mis mes rideaux de bleu et blanc, je me suis rendu compte de la tige de métal noir mat était tordu et didn't avec le reste de la salle.
✗	CNN	**Comme** je mets mes rideaux bleus et blancs, j' ai réalisé la tige en métal noir matte The staff were very friendly and helpful.
✓	RNN	**Lorsque** j' ai mis mes rideaux bleus et blancs, j' ai réalisé que la tige de métal noir matte était collée et ne correspondait pas avec le reste de la pièce.
✓	Attention	**Au moment où** j' ai mis mes rideaux bleus et blancs, je me suis rendu compte que la tige de métal noir matte était crookée et ne correspondait pas au reste de la pièce.
✓	Google	**En remontant** mes rideaux bleus et blancs, j'ai réalisé que la tringle en métal noir mat était tordue et ne correspondait pas au reste de la pièce.
✓	DeepL	**Alors que** je montais mes rideaux bleu et blanc, je me suis rendu compte que la tige métallique noire mate était tordue et ne correspondait pas au reste de la pièce.

"as" between the propositions		
S1g	Source	My grandmother laughed at me and told me I was getting old when my knees cracked **as** I stood up to go grab my phone.
	Ref	Ma grand-mère s'est moquée de moi et m'a dit que je vieillissais quand mes genoux ont craqué **alors que** je me levais pour aller chercher mon téléphone.
✗	Hybrid SMT	Ma grand-mère m' a RI et m' a dit que je recevais lorsqu' il a craqué mes genoux **comme** je me suis levé pour aller prendre mon téléphone.
✗	CNN	Ma grand-mère m' a ri et m' a dit que j' étais vieux lorsque mes genoux craquaient **comme** je l' ai laissé s' approprier mon téléphone.
✓	RNN	Ma grand-mère m' a rêvé et m' a dit que j' étais vieux lorsque mes genoux se fissuraient **alors que** j' avais pris mon téléphone.
✗*	Attention	Ma grand-mère m' a ri et m' a dit que j' avais vieilli lorsque mes genoux se sont fissurés pour prendre mon téléphone.
✓	Google	Ma grand-mère a ri de moi et m'a dit que je vieillissais quand mes genoux se sont fissurés **alors que** je me levais pour aller attraper mon téléphone.
✓	DeepL	Ma grand-mère s'est moquée de moi et m'a dit que je devenais vieux quand mes genoux ont craqué **alors que** je me levais pour aller chercher mon téléphone.
S1h	Source	I discovered a hidden image that seemed to be one of a sheep **as** I was restoring an old painting from the sixteenth century, depicting a biblical scene.
	Ref	J'ai découvert une image cachée qui semblait être celle d'un mouton **alors que** je restaurais une ancienne peinture du seizième siècle qui représentait une scène
✗	Hybrid SMT	J' ai découvert une image cachée qui semblait être l' un des moutons **comme** j' étais restaurer une ancienne peinture du seizième siècle, illustrant une scène biblique.
✗	CNN	J' ai découvert une image cachée qui semblait être l' un des moutons **que** j' étais en train de restaurer un vieux peinture du XVIe siècle, montrant une scène biblique.
✓	RNN	J' ai découvert une image cachée qui semblait être l' une d' un mouton **alors que** je restaurais une vieille peinture datant du XVIe siècle, représentant une scène biblique.

✓	Attention	J' ai découvert une image cachée qui semblait être l' un des moutons **alors que** je restaurais une vieille peinture du XVIe siècle, représentant une scène biblique.
✓	Google	J'ai découvert une image cachée qui semblait être celle d'un mouton **alors que** je restaurais un vieux tableau du XVIe siècle, représentant une scène biblique.
✓	DeepL	J'ai découvert une image cachée qui semblait être celle d'un mouton **alors que** je restaurais un vieux tableau du XVIe siècle, représentant une scène biblique.

"As" expressing a cause

Possible solutions: *puisque*, *parce que*, *car*, etc.

Short sentences		
"as" before the propositions		
S2a	Source	**As** I kept being woken up by raccoons, I put traps in my backyard.
	Ref	**Comme** je me faisais toujours réveiller par des ratons laveurs, j'ai placé des pièges dans ma cour.
✓	Hybrid SMT	**Comme** je l' ai gardé d' être réveillé par les ratons laveurs, je pose des pièges dans mon arrière-cour.
✗	CNN	**Alors que** je me suis réveillé par des ratons laveurs, j ' ai mis des pièges dans mes l' arrière-cour.
✓	RNN	**Comme** je me suis toujours réveillé par les ratons laveurs, j' ai mis des pièges dans mon arrière-cour.
✗	Attention	**Au fur et à mesure que** j' ai continué d' être soulevée par des raccoons, j' ai mis des pièges dans mon arrière-cour.
✗	Google	**Alors que** je continuais à être réveillé par des ratons laveurs, j'ai posé des pièges dans mon jardin.
✓	DeepL	**Comme** les ratons laveurs ne cessaient de me réveiller, je posais des pièges dans mon jardin.
S2b	Source	**As** I spoke without raising my hand, the teacher ignored my answer.
	Ref	**Comme** j'ai parlé sans lever la main, le professeur a ignoré ma réponse.
✓	Hybrid SMT	**Comme** je parlais sans soulever ma main, l' enseignant a ignoré ma réponse.
✓	CNN	**Comme** je l' ai parlé sans m' élever, l' enseignant a ignoré ma réponse.

✓	RNN	**Comme** je l' ai parlé sans élever la main, l' enseignant a ignoré ma réponse.
✓	Attention	**Comme** je l' ai parlé sans soulever la main, l' enseignant a ignoré ma réponse.
✗	Google	**Alors que** je parlais sans lever la main, le professeur ignora ma réponse.
✓	DeepL	**Comme** je parlais sans lever la main, le professeur a ignoré ma réponse.
"as" between the propositions		
S2c	Source	The doctor reduced the dosage, **as** the patient seemed to experience discomfort.
	Ref	Le médecin a réduit la posologie **car** le patient semblait ressentir un inconfort.
✓	Hybrid SMT	Le médecin a réduit la dose, **comme** le patient semblait faire l' expérience de l' inconfort.
✓	CNN	Le médecin a réduit la dose, **car** le patient semblait éprouver de l' inconfort.
✓	RNN	Le médecin a réduit la dose, **car** le patient semblait souffrir d' inconfort.
✓	Attention	Le médecin a réduit la dose, **car** le patient semblait souffrir d' inconfort.
✓	Google	Le médecin a réduit la posologie, **car** le patient semblait ressentir une gêne.
✓	DeepL	Le médecin a réduit la dose, **car** le patient semblait éprouver un certain malaise.
S2d	Source	Hikers tend to avoid that trail, **as** the path is narrow.
	Ref	Les randonneurs ont tendance à éviter ce chemin **car** le sentier est étroit.
✓	Hybrid SMT	Les randonneurs ont tendance à éviter que le sentier, **comme** la voie est étroite.
✓	CNN	Les randonneurs ont tendance à éviter ce sentier, **car** le sentier est étroit.
✓	RNN	Les randonneurs ont tendance à éviter ce sentier, **car** le sentier est étroit.
✓	Attention	Les randonneurs ont tendance à éviter ce sentier, **car** le sentier est étroit.
✓	Google	Les randonneurs ont tendance à éviter ce sentier, **car** le chemin est étroit.
✓	DeepL	Les randonneurs ont tendance à éviter ce sentier, **car** le chemin est étroit.

Long sentences		
"as" before the propositions		
S2e	Source	**As** I kept being woken up by raccoons opening my garbage bins every night, I put traps with peanut butter and jelly in my backyard.
	Ref	**Comme** je me faisais toujours réveiller par des ratons laveurs qui ouvraient mes poubelles chaque nuit, j'ai placé des pièges avec du beurre d'arachide et de la confiture dans ma cour.
✓	Hybrid SMT	**Comme** je l' ai gardé d' être réveillé par les ratons-laveurs ouvrir mes poubelles chaque soir, j' ai mis les pièges avec le beurre d' arachide et la confiture dans mon arrière-cour.
✓	CNN	**Comme** je l' ai réveillé par des raccoons ouvrant mes poubelles tous les soirs, j' ai mis des pièges avec du beurre d' arachide et de gelée dans ma cour.
✗	RNN	**Lorsque** je me suis rendu compte que les ratons laveurs ouvriraient mes poubelles chaque soir, je mets des pièges avec du beurre d ' arachide et de la gelée dans ma cour.
✗	Attention	**Alors que** j' ai continué d' être soulevée par des raccoons ouvrant mes poubelles tous les soirs, j' ai mis des pièges avec du beurre d' arachides.
✗	Google	**Alors que** je n'arrêtais pas d'être réveillé par des ratons laveurs qui ouvraient mes poubelles tous les soirs, j'ai mis des pièges avec du beurre d'arachide et de la gelée dans mon jardin.
✓	DeepL	**Comme** je continuais à être réveillé par des ratons laveurs qui ouvraient mes poubelles chaque nuit, j'ai mis des pièges avec du beurre de cacahuète et de la gelée dans mon jardin.
S2f	Source	**As** I spoke in class without raising my hand and waiting to be called on, the geography teacher ignored my answer and asked someone else to respond.
	Ref	**Comme** j'ai parlé en classe sans lever la main et sans attendre d'avoir la parole, le professeur de géographie a ignoré ma réponse et a demandé à quelqu'un d'autre de répondre.
✓	Hybrid SMT	**Comme** j' ai parlé de la classe sans soulever ma main et attendent d' être appelée, l' enseignant de géographie a ignoré ma réponse et a demandé à quelqu' un d' autre pour répondre.

✓	CNN	**Comme** je l' ai parlé en classe sans m' élever la main et attendre d' être appelée, l' enseignant de la géographie ignoré ma réponse et demandait à quelqu' un d' autre de répondre.
✓	RNN	**Comme** j' ai parlé en classe sans m' élever et en attendant d' être appelé, le professeur de géographie a ignoré ma réponse et demandé à quelqu' un d' autre de répondre.
✓	Attention	**Comme** j' ai parlé en classe sans élever la main et attendre d' être appelé, l' enseignant en géographie a ignoré ma réponse et a demandé à quelqu' un d' autre de répondre.
✗	Google	**Alors que** je parlais en classe sans lever la main et attendre d'être appelé, le professeur de géographie a ignoré ma réponse et a demandé à quelqu'un d'autre de répondre.
✗	DeepL	**Alors que** je parlais en classe sans lever la main et en attendant qu'on m'appelle, le professeur de géographie a ignoré ma réponse et a demandé à quelqu'un d'autre de répondre.
"as" between the propositions		
S2g	Source	The doctor reduced the dosage of the medication he had prescribed two weeks ago, **as** the young patient had experienced discomfort ever since he started taking those pills.
	Ref	Le médecin a réduit la posologie du médicament qu'il avait prescrit voilà deux semaines **car** le jeune patient semblait ressentir un inconfort depuis qu'il avait commencé à prendre ces pilules.
✗	Hybrid SMT	Le médecin a réduit la dose des médicaments qu' il avait prescrit il y a deux semaines, **en tant que** jeune patient avait connu l' inconfort depuis qu' il a commencé à prendre ces pilules.
✗	CNN	Le médecin a réduit le dosage du médicament qu' il avait prescrit il y a deux semaines, **alors que** le jeune patient n' avait jamais eu de malaise depuis qu' il a commencé à prendre ces pilules.
✓	RNN	Le médecin a réduit la dose de médicament qu' il avait prescrit il y a deux semaines, **car** le jeune patient avait connu un malaise depuis qu' il a commencé à prendre ces pilules.

✓	Attention	Le médecin a réduit la posologie du médicament qu' il avait prescrit il y a deux semaines, **puisque** le jeune patient avait eu de l' inconfort depuis qu' il avait commencé à prendre ces pilules.
✓	Google	Le médecin a réduit la posologie du médicament qu'il avait prescrit il y a deux semaines, **car** le jeune patient éprouvait une gêne depuis qu'il avait commencé à prendre ces pilules.
✓	DeepL	Le médecin a réduit le dosage des médicaments qu'il avait prescrits il y a deux semaines, **car** le jeune patient ressentait un malaise depuis qu'il avait commencé à prendre ces pilules.
S2h	Source	Hikers tend to avoid that trail on the west side of the park near the gates, **as** the path is narrow and is usually hard to follow.
	Ref	Les randonneurs ont tendance à éviter ce chemin sur le côté ouest du parc, près des barrières, **car** le sentier est étroit et il est difficile à suivre.
✓	Hybrid SMT	Les randonneurs ont tendance à éviter que le sentier du côté Ouest du parc, près de la porte, **comme** la voie est étroite et est habituellement difficile à suivre.
✗	CNN	Les randonneurs ont tendance à éviter ce sentier du côté ouest du parc près des portes, **comme suit**: Le chemin est étroit et est habituellement difficile à suivre.
✓	RNN	Les randonneurs ont tendance à éviter ce sentier du côté ouest du parc près des portes, **car** le sentier est étroit et est habituellement difficile à suivre.
✓	Attention	Les randonneurs ont tendance à éviter ce sentier du côté ouest du parc près des portes, **car** le sentier est étroit et est habituellement difficile à suivre.
✓	Google	Les randonneurs ont tendance à éviter ce sentier du côté ouest du parc près des portes, **car** le chemin est étroit et généralement difficile à suivre.
✓	DeepL	Les randonneurs ont tendance à éviter ce sentier sur le côté ouest du parc, près des portes, **car** le chemin est étroit et généralement difficile à suivre.

"As" expressing progression

Possible solutions: *au fur et à mesure que*, *à mesure que*, etc.

Short sentences		
"as" before the propositions		
S3a	Source	**As** this project develops, we'll have to hire more employees.
	Ref	**À mesure que** ce projet évoluera, nous devrons embaucher plus d'employés.
✗	Hybrid SMT	**Comme** ce projet développe, we'll à embaucher plus d' employés.
✗	CNN	**Étant donné que** ce projet se développe, nous devrons embaucher plus d' employés.
✓	RNN	**Au fur et à mesure de** l' élaboration de ce projet, nous devrons embaucher plus d' employés.
✓	Attention	**Au fur et à mesure que** ce projet se développera, nous devrons embaucher davantage d' employés.
✓	Google	**À mesure que** ce projet se développe, nous devrons embaucher plus d'employés.
✓	DeepL	**Au fur et à mesure que** ce projet se développera, nous devrons embaucher davantage de salariés.
S3b	Source	**As** we receive the comments, we'll compile them in one document.
	Ref	**À mesure que** nous recevrons les commentaires, nous les compilerons en un document.
✗	Hybrid SMT	**Comme** nous recevons les commentaires, we'll les compiler dans un seul document.
✓	CNN	**À mesure que** nous recevons les commentaires, nous les compilerons dans un seul document.
✓	RNN	**Au fur et à mesure que** nous recevrons les commentaires, nous les compilerons dans un seul document.
✓	Attention	**Au fur et à mesure que** nous recevons les commentaires, nous les compilerons dans un seul document.
✓	Google	**Au fur et à mesure que** nous recevrons les commentaires, nous les compilerons dans un seul document.
✓	DeepL	**Au fur et à mesure que** nous recevrons les commentaires, nous les rassemblerons dans un document unique.
"as" between the propositions		
S3c	Source	We will move to a bigger house **as** our family grows.
	Ref	Nous allons déménager dans une plus grande maison **à mesure que** notre famille s'agrandira.

✗	Hybrid SMT	Nous adopterons une maison plus grande **que** notre famille se développe.
✗	CNN	Nous passerons à une plus grande maison **que** notre famille s' accroît.
✓	RNN	Nous allons passer à une maison plus grande **au fur et à mesure que** notre famille croît.
✓	Attention	Nous déménagerons dans une maison plus grande **au fur et à mesure que** notre famille grandit.
✓	Google	Nous déménagerons dans une maison plus grande **à mesure que** notre famille s'agrandira.
✓	DeepL	Nous déménagerons dans une maison plus grande **au fur et à mesure que** notre famille s'agrandira.
S3d	Source	To make risotto, you must add small quantities of broth as the rice cooks.
	Ref	Pour faire un risotto, vous devez ajouter le bouillon en petites quantités **à mesure que** le riz cuit.
✗	Hybrid SMT	Pour rendre le risotto, vous devez ajouter de petites quantités de bouillon **comme** les cuisiniers de riz.
✗*	CNN	Pour faire de la risotto, vous devez ajouter de petites quantités de bouillon **que** les cuisiniers de riz.
✗	RNN	Pour risotto, vous devez ajouter de petites quantités de bouillon **comme** les cuisiniers.
✗	Attention	Pour faire de la risotto, vous devez ajouter de petites quantités de bouillon **comme** cuisinières de riz.
✓	Google	Pour faire un risotto, vous devez ajouter de petites quantités de bouillon **pendant** la cuisson du riz.
✓	DeepL	Pour faire du risotto, vous devez ajouter de petites quantités de bouillon **au fur et à mesure que** le riz cuit.
Long sentences		
"as" before the propositions		
S3e	Source	**As** this new project develops, it will require expertise from experienced workers, and we'll have to hire more full-time employees with the required language proficiency.
	Ref	**À mesure que** ce nouveau projet évoluera, il demandera l'expertise des travailleurs d'expérience et nous devrons embaucher plus d'employés à temps plein ayant le niveau de compétence linguistique requis.

✗	Hybrid SMT	**Comme** ce nouveau projet se développe, il faudra que le savoir-faire de travailleurs expérimentés, et we'll à embaucher plus d' employés à temps plein avec les compétences linguistiques requises.
✗	CNN	**Étant donné que** ce nouveau projet se développe, il aura besoin de l' expertise de travailleurs expérimentés, et nous y sommes. Il faut embaucher davantage d' employés à temps plein possédant la compétence linguistique requise.
✓	RNN	**Au fur et à mesure de** l' évolution de ce nouveau projet, il faudra l' expertise des travailleurs expérimentés, et nous devrons embaucher plus d' employés à temps plein possédant les compétences linguistiques requises.
✓	Attention	**Au fur et à mesure que** ce nouveau projet se développera, il exigera l' expertise de travailleurs expérimentés et nous devrons embaucher plus d' employés à temps plein possédant les compétences linguistiques requises.
✓	Google	**À mesure que** ce nouveau projet se développera, il exigera l'expertise de travailleurs expérimentés et nous devrons embaucher davantage d'employés à temps plein possédant les compétences linguistiques requises.
✓	DeepL	**Au fur et à mesure que** ce nouveau projet se développera, il nécessitera l'expertise de travailleurs expérimentés, et nous devrons embaucher davantage d'employés à plein temps ayant les compétences linguistiques requises.
S3f	Source	**As** we receive the comments from the people who participated in our workshop last week, we'll compile them in one document and share the feedback.
	Ref	**À mesure que** nous recevrons les commentaires des personnes ayant participé à notre atelier la semaine dernière, nous les compilerons en un document et nous partagerons la rétroaction.
✗	Hybrid SMT	**Comme** nous recevons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, we'll les compiler dans un seul document et de partager les commentaires.
✗	CNN	**Alors que** nous recevons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous sommes Les compiler dans un seul document et partager les commentaires.

✓	RNN	**Au fur et à mesure que** nous recevrons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous les compilerons dans un seul document et partagerons les commentaires.
✓	Attention	**À mesure que** nous recevrons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous les compilerons dans un seul document et partagerons les commentaires.
✓	Google	**Au fur et à mesure que** nous recevons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous les compilerons dans un seul document et partagerons les commentaires.
✓	DeepL	**Au fur et à mesure que** nous recevrons les commentaires des personnes qui ont participé à notre atelier la semaine dernière, nous les rassemblerons dans un document et partagerons les réactions.
"as" between the propositions		
S3g	Source	We will move to a bigger house in a nicer neighbourhood, closer to our work, **as** our family grows and we need more bedrooms for the kids.
	Ref	Nous allons déménager dans une plus grande maison, dans un plus beau quartier, plus proche de notre travail, **à mesure que** notre famille s'agrandira et que nous aurons besoin de plus de chambres pour les enfants.
✗	Hybrid SMT	Nous adopterons une maison plus grande dans un quartier plus agréable, plus proche de notre travail, **comme** notre famille croît et nous avons besoin de plus de chambres pour les enfants.
✗	CNN	Nous passerons à une plus grande maison dans un quartier plus agréable, plus près de notre travail, **à savoir:** Notre famille grandit et nous avons besoin de plus de chambres pour les enfants.
✓	RNN	Nous allons passer à une maison plus grande dans un quartier plus agréable, plus près de notre travail, **au fur et à mesure que** notre famille grandit et que nous avons besoin de plus de chambres pour les enfants.
✓	Attention	Nous déménagerons dans une maison plus grande dans un quartier plus agréable, plus proche de notre travail, **à mesure que** notre famille grandit et que nous avons besoin de plus de chambres pour les enfants.

✗	Google	Nous déménagerons dans une maison plus grande dans un quartier plus agréable, plus proche de notre travail, **car** notre famille s'agrandit et nous avons besoin de plus de chambres pour les enfants.
✓	DeepL	Nous allons déménager dans une maison plus grande, dans un quartier plus agréable, plus proche de notre travail, **à mesure que** notre famille s'agrandit et que nous avons besoin de plus de chambres pour les enfants.
S3h	Source	According to this recipe I found online, to make risotto, you must add the chicken or vegetable broth in small quantities **as** the rice cooks over medium heat.
	Ref	D'après cette recette que j'ai trouvée en ligne, pour faire un risotto, vous devez ajouter le bouillon de poulet ou de légumes en petite quantité **à mesure que** le riz cuit à température moyenne.
✗	Hybrid SMT	Selon cette recette, j' ai trouvé en ligne, pour rendre le risotto, vous devez ajouter le poulet ou le bouillon de légumes en petites quantités **comme** le riz cuit à feu moyen.
✗	CNN	Selon cette recette, j' ai trouvé en ligne, pour faire risotto, vous devez ajouter le poulet ou de bouillon de légumes en petites quantités **que** les cuisiniers à feu moyen.
✗	RNN	Selon cette recette que j' ai trouvée en ligne, pour faire du risotto, vous devez ajouter le poulet ou le bouillon de légumes en petites quantités **lorsque** le riz cuira à feu moyen.
✓	Attention	Selon cette recette que j' ai trouvée en ligne, pour faire de la risotto, vous devez ajouter le bouillon de poulet ou de légumes en petites quantités **à mesure que** le riz cuisine.
✓	Google	Selon cette recette que j'ai trouvée en ligne, pour faire du risotto, vous devez ajouter le bouillon de poulet ou de légumes en petites quantités **pendant que** le riz cuit à feu moyen.
✓	DeepL	Selon cette recette que j'ai trouvée en ligne, pour faire un risotto, il faut ajouter le bouillon de poulet ou de légumes en petites quantités **au fur et à mesure que** le riz cuit à feu moyen.

"While" expressing temporality

Possible solutions: *en*, *pendant que*, *tout en*, etc.

Short sentences		
"while" before the propositions		
S4a	Source	**While** I was cleaning my room, I noticed a letter I had never opened.
	Ref	**Pendant que** je nettoyais ma chambre, j'ai remarqué une lettre que je n'avais jamais ouverte.
✗	Hybrid SMT	**Alors que** j' étais nettoyer ma chambre, j' ai remarqué une lettre que je n' avais jamais ouvert.
✓*	CNN	J' étais en train de nettoyer ma chambre, j' ai remarqué une lettre que je n' avais jamais ouvert.
✓	RNN	**Pendant que** je nettoyais ma chambre, j' ai remarqué une lettre que je n' avais jamais ouverte.
✓	Attention	**Pendant que** j' ai nettoyé ma chambre, j' ai remarqué une lettre que je n' avais jamais ouverte.
✓	Google	**Pendant que** je nettoyais ma chambre, j'ai remarqué une lettre que je n'avais jamais ouverte.
✗	DeepL	**Alors que** je nettoyais ma chambre, j'ai remarqué une lettre que je n'avais jamais ouverte.
S4b	Source	**While** I was on the phone, the deliveryman rang the doorbell.
	Ref	**Pendant que** j'étais au téléphone, le livreur a sonné à la porte.
✗	Hybrid SMT	**Alors que** j' étais au téléphone, le livreur a sonné.
✓	CNN	**Pendant que** j' étais au téléphone, le responsable de l' exécuteur allait la porte.
✓	RNN	**Pendant que** j' étais au téléphone, l' homme de la livraison s' occupait de la porte de la porte.
✓	Attention	**Pendant que** j' étais au téléphone, l' expéditeur a rangé la porte.
✓	Google	**Pendant que** j'étais au téléphone, le livreur a sonné à la porte.
✗	DeepL	**Alors que** j'étais au téléphone, le livreur a sonné à la porte.
"while" between the propositions		
S4c	Source	He bought new glasses **while** he was visiting.
	Ref	Il a acheté de nouvelles lunettes **lors de** sa visite.
✗	Hybrid SMT	Il a acheté de nouvelles lunettes, **alors qu' il** était en visite.

✓	CNN	Il a acheté de nouveaux verres **pendant** sa visite.
✓	RNN	Il a acheté de nouveaux lunettes **pendant** qu' il visitait.
✓	Attention	Il a acheté de nouvelles lunettes **lors de** sa visite.
✓	Google	Il a acheté de nouvelles lunettes **lors de** sa visite.
✓	DeepL	Il a acheté de nouvelles lunettes **pendant** sa visite.
S4d	Source	He fell asleep **while** reading her eighteen-page-long letter.
	Ref	Il s'est endormi **en lisant** sa lettre de dix-huit pages.
✗	Hybrid SMT	Il s' endormit **tandis que** la lecture de sa lettre eighteen-page-long.
✗	CNN	Il s' endormait **alors qu'** elle lisait sa lettre de dix-huit pages.
✓	RNN	Il s' est endormi **en lisant sa** lettre de dix-huit pages.
✓	Attention	Il est tombé aslepte **en lisant** sa lettre d' une durée de dix-huit pages.
✓	Google	Il s'est endormi **en lisant** sa lettre de dix-huit pages.
✓	DeepL	Il s'est endormi **en lisant** sa lettre de dix-huit pages.
Long sentences		
"while" before the propositions		
S4e	Source	**While** I was cleaning my old room at my parents' place, I noticed a letter I had never opened from a childhood friend with whom I had not talked in years.
	Ref	**Pendant que** je nettoyais mon ancienne chambre chez mes parents, j'ai remarqué une lettre que je n'avais jamais ouverte, de la part de mon ami d'enfance à qui je n'ai pas parlé depuis des années.
✗	Hybrid SMT	**Alors que** j' étais nettoyer ma salle de vieux à mes parents' place, j' ai remarqué une lettre que je n' avais jamais ouvert à partir d' un ami d' enfance avec qui je n' avais pas parlé depuis des années.
✗	CNN	**Même si** j' étais en train de nettoyer ma vieille pièce à l' endroit de mes parents, j' ai remarqué une lettre Je n' avais jamais ouvert d' un ami d' enfance avec qui je n' avais pas parlé depuis des années.
✓	RNN	**Pendant que** je nettoyais ma vieille chambre au lieu de mes parents, j' ai remarqué une lettre que je n' avais jamais ouverte d' un ami avec lequel je n' avais pas parlé depuis des années.

✓	Attention	**Pendant que** j' ai nettoyé ma vieille chambre à l' endroit de mes parents, j' ai remarqué une lettre que je n' avais jamais ouverte d' un ami d' enfance.
✓	Google	**Pendant que** je nettoyais mon ancienne chambre chez mes parents, j'ai remarqué une lettre que je n'avais jamais ouverte d'un ami d'enfance avec qui je n'avais pas parlé depuis des années.
✗	DeepL	**Alors que** je nettoyais mon ancienne chambre chez mes parents, j'ai remarqué une lettre que je n'avais jamais ouverte et qui provenait d'un ami d'enfance avec qui je n'avais pas parlé depuis des années.
S4f	Source	**While** I was on the phone discussing a COVID-19 response project with three of my colleagues, the deliveryman rang the doorbell, carrying three stacked boxes.
	Ref	**Pendant que** je discutais d'un projet d'intervention concernant la COVID-19 avec trois de mes collègues au téléphone, le livreur a sonné à la porte avec trois boîtes empilées dans les bras.
✗	Hybrid SMT	**Alors que** j' étais au téléphone à discuter d' un projet de réponse COVID-19 avec trois de mes collègues, le livreur a sonné, transporter trois boîtes empilées.
✓	CNN	**Pendant que** j' étais au téléphone pour discuter d' un projet d' intervention COVID-19 avec trois de mes collègues, le responsable de l' exécuteur allait la porte, transportant trois boîtes empilées.
✓	RNN	**Pendant que** j ' étais au téléphone pour discuter d ' un projet d ' intervention COVID-19 avec trois de mes collègues, le chargé de livraison s ' est occupé de trois boîtes empilées.
✗	Attention	**Alors que** j' étais au téléphone pour discuter d' un projet d' intervention COVID-19 avec trois de mes collègues, l' expéditeur a rangé la porte, portant trois boîtes empilées.
✗	Google	**Alors que** j'étais au téléphone pour discuter d'un projet de réponse au COVID-19 avec trois de mes collègues, le livreur a sonné à la porte, transportant trois boîtes empilées.
✗	DeepL	**Alors que** j'étais au téléphone en train de discuter d'un projet de réponse COVID-19 avec trois de mes collègues, le livreur a sonné à la porte, transportant trois boîtes empilées.

"while" between the propositions		
S4g	Source	He bought a new pair of reading glasses that, in my opinion, are too big for his face, **while** he was visiting his mom in New Hampshire.
	Ref	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, est trop grande pour son visage, **pendant qu'**il rendait visite à sa mère au New Hampshire.
✗	Hybrid SMT	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grandes pour son visage, **alors qu'** il visitait sa mère dans le New Hampshire.
✗	CNN	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grandes pour s' élever. son visage, **alors qu '** il visitait sa mère au New Hampshire.
✗	RNN	Il achète une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grandes pour son visage, **alors qu' il** visitait sa mère au New Hampshire.
✗	Attention	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grosses pour son visage, **alors qu'** il visitait sa mère au New Hampshire.
✗	Google	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grandes pour son visage, **alors qu'**il rendait visite à sa mère dans le New Hampshire.
✗	DeepL	Il a acheté une nouvelle paire de lunettes de lecture qui, à mon avis, sont trop grandes pour son visage, **alors qu'**il rendait visite à sa mère dans le New Hampshire.
S4h	Source	He fell asleep at the white kitchen table that he had not cleaned up after last night's dinner, and that still held a dirty plate and glass, **while** reading her eighteen-page-long letter.
	Ref	Il s'est endormi à la table de cuisine blanche qu'il n'avait pas encore lavée après le souper d'hier soir, et sur laquelle se trouvait encore une assiette sale et une verre, **en lisant** sa lettre de dix-huit pages.
✗	Hybrid SMT	Il s' endormit à la table de la cuisine blanche qu' il n' avait pas été décontaminée après dîner 's hier soir, et qui détiennent encore une plaque sale et le verre, **tandis que** la lecture de sa lettre eighteen-page-long.
✗	CNN	Il s' est endormi à la table de cuisine blanche qu' il n' avait pas nettoyé après la nuit passée Le souper, et qui détenait toujours une plaque et un verre sale, **tout en la lisant** dix-huit Lettre d' une page.

✗	RNN	Il s' est endormi à la table blanche de cuisine qu' il n' avait pas nettoyé après le dîner de nuit, et qui tenait encore une plaque sale et du verre, **tout en lisant** sa lettre de 18 pages.
✗	Attention	Il est tombé à la table de cuisine blanche qu' il n' avait pas nettoyé après le dîner de la nuit dernière et qu' il gardait encore une plaque et un verre sales, **tout en lisant** sa lettre de 18 pages.
✓	Google	Il s'est endormi à la table blanche de la cuisine qu'il n'avait pas nettoyée après le dîner d'hier soir, et qui contenait encore une assiette et un verre sales, **en lisant** sa lettre de dix-huit pages.
✗	DeepL	Il s'est endormi sur la table blanche de la cuisine qu'il n'avait pas nettoyée après le dîner d'hier soir, et qui contenait encore une assiette et un verre sales, **tout en lisant** sa lettre de dix-huit pages.

"While" expressing a concession

Possible solutions: *bien que*, *malgré que*, *même si*, *quoique*, or *si*

Short sentences		
"while" at the beginning of the sentence		
S5a	Source	**While** I enjoy your company, I'm going to have to ask you to leave.
	Ref	**Bien que** j'apprécie votre compagnie, je vais devoir vous demander de partir.
✓	Hybrid SMT	**Bien que** j' aime votre compagnie, I'm vont devoir vous demander de quitter.
✓	CNN	**Bien que** j' apprécie votre entreprise, je vais devoir vous demander de partir.
✗	RNN	**Pendant que** j' aime votre entreprise, je vais vous demander de partir.
✗	Attention	**Pendant que** j' aime votre entreprise, je vais devoir vous demander de partir.
✗	Google	**Tant que** j'apprécie votre compagnie, je vais devoir vous demander de partir.
✓	DeepL	**Bien que** j'apprécie votre compagnie, je vais devoir vous demander de partir.

S5b	Source	**While** money is always tight, I can generally make a living freelancing.
	Ref	**Bien que** l'argent se fasse rare, je peux généralement gagner ma vie en faisant du travail à la pige.
✓	Hybrid SMT	**Bien que** l' argent est toujours serré, je peux généralement vivre de la pige.
✓	CNN	**Bien que** l' argent soit toujours serré, je peux généralement faire un freelancing vivant.
✓	RNN	**Bien que** l' argent soit toujours très serré, je peux généralement faire un freelancing vivant.
✓	Attention	**Bien que** l' argent soit toujours serré, je peux généralement faire un freelancing vivant.
✓	Google	**Bien que** l'argent soit toujours serré, je peux généralement gagner ma vie en freelance.
✓	DeepL	**Bien que** l'argent soit toujours rare, je peux généralement gagner ma vie en travaillant en free-lance.
"while" mid-sentence		
S5c	Source	The actor, **while** handsome and charming, was not a particularly nice person to work with.
	Ref	L'acteur, **bien que** beau et charmant, n'était pas une personne particulièrement agréable avec qui travailler.
✗	Hybrid SMT	L' acteur, **tout** beau et charmant, n' était pas particulièrement bonne personne pour travailler.
✓	CNN	L' acteur, **bien qu'** agréable et charmant, n' était pas une personne particulièrement agréable pour travailler avec.
✗	RNN	L' acteur, **alors qu'** il était beau et charmant, n' était pas particulièrement agréable à travailler avec.
✓	Attention	L' acteur, **bien que** gentil et charmant, n' était pas une personne particulièrement agréable à travailler.
✓	Google	L'acteur, **bien que** beau et charmant, n'était pas une personne particulièrement agréable à travailler.
✓	DeepL	L'acteur, **bien que** beau et charmant, n'était pas une personne particulièrement agréable avec qui travailler.
S5d	Source	The hotel room, **while** it was clean and cheap, was not particularly attractive.
	Ref	La chambre d'hôtel, **bien que** propre et abordable, n'était pas particulièrement attrayante.
✗	Hybrid SMT	La chambre d' hôtel, **alors qu'** elle était propre et bon marché, n' a pas été particulièrement attrayante.

✓	CNN	L' hôtel, **bien qu'** il soit propre et bon marché, n' était pas particulièrement attrayant.
✗*	RNN	L' hôtel est très bien situé, à proximité de l' aéroport.
✗*	Attention	L' hôtel est très bien situé, à proximité de la gare, de l' aéroport, de la gare, de la gare et de la gare.
✓	Google	La chambre d'hôtel, **bien qu**'elle soit propre et bon marché, n'était pas particulièrement attrayante.
✓	DeepL	La chambre d'hôtel, **bien que** propre et bon marché, n'était pas particulièrement attirante.
Long sentences		
"while" at the beginning of the sentence		
S5e	Source	**While** I enjoy your company and I appreciate your help, I'm going to have to ask you to leave and take your dog with you.
	Ref	**Bien que** j'apprécie votre compagnie et votre aide, je vais devoir vous demander de partir et d'emmener votre chien avec vous.
✓	Hybrid SMT	**Bien que** j' aime votre compagnie et j' apprécie votre aide, I'm vont devoir vous demander de quitter et prendre votre chien avec vous.
✓	CNN	**Bien que** j' apprécie votre entreprise et j' apprécie votre aide, je vais avoir vous demandez de partir et de prendre votre chien avec vous.
✗	RNN	**Pendant que** j' apprécie votre entreprise et que j' apprécie votre aide, je vais vous demander de partir et de prendre votre chien avec vous.
✗	Attention	**Pendant que** j' apprécie votre entreprise et que j' apprécie votre aide, je vais devoir vous demander de partir et de prendre votre chien avec vous.
✓	Google	**Bien que** j'apprécie votre compagnie et que j'apprécie votre aide, je vais devoir vous demander de partir et d'emmener votre chien avec vous.
✓	DeepL	**Bien que** j'apprécie votre compagnie et votre aide, je vais devoir vous demander de partir et de prendre votre chien avec vous.

S5f	Source	**While** money is always tight and rent for my small apartment is only getting more and more expensive, I can generally make a living freelancing.
	Ref	**Bien que** l'argent se fasse rare et que le loyer pour mon petit appartement ne cesse d'augmenter, je peux généralement gagner ma vie en faisant du travail à la pige.
✓	Hybrid SMT	**Bien que** l' argent est toujours serré et louer pour mon petit appartement n' est que de plus en plus cher, je peux généralement vivre de la pige.
✓	CNN	**Même si** l' argent est toujours serré et le loyer pour mon petit appartement n' est que de plus en plus cher, I peut gÃ © nÃ © ralement un freelancing vivant.
✓	RNN	**Bien que** l' argent soit toujours serré et le loyer pour mon petit appartement est de plus en plus coûteux, je peux généralement faire un freelancing vivant.
✓	Attention	**Bien que** l' argent soit toujours serré et que le loyer pour mon petit appartement ne soit que de plus en plus coûteux, je peux généralement faire un freelancing vivant.
✓	Google	**Bien que** l'argent soit toujours serré et que le loyer de mon petit appartement ne devienne que de plus en plus cher, je peux généralement gagner ma vie à la pige.
✓	DeepL	**Bien que** l'argent soit toujours rare et que le loyer de mon petit appartement ne cesse de grimper, je peux généralement gagner ma vie en travaillant en free-lance.
"while" mid-sentence		
S5g	Source	The actor, **while** handsome and charming and selected among hundreds of others for the upcoming science fiction movie, was not a particularly nice person to work with.
	Ref	L'acteur, **bien que** beau, charmant et sélectionné parmi des centaines d'autres acteurs pour le prochain film de science fiction, n'était pas une personne particulièrement agréable avec qui travailler.
✗	Hybrid SMT	L' acteur, **tout** beau et charmant et sélectionnés parmi des centaines d' autres pour le prochain film de science-fiction, n' était pas particulièrement bonne personne pour travailler.
✓	CNN	L' acteur, **bien qu'** il soit agréable et charmant et choisi parmi des centaines d' autres pour la science à venir Le film de fiction, n' était pas une personne particulièrement agréable pour travailler avec.

✓	RNN	L' acteur, **si** beau et charmant et choisi parmi des centaines d' autres pour le prochain film de science-fiction, n' était pas une personne particulièrement agréable à travailler.
✓	Attention	L' acteur, **bien que** gentil et charmant, choisi parmi des centaines d' autres pour le prochain film de science-fiction, n' était pas une personne particulièrement agréable à travailler.
✓	Google	L'acteur, **bien que** beau et charmant et sélectionné parmi des centaines d'autres pour le prochain film de science-fiction, n'était pas une personne particulièrement agréable à travailler.
✓	DeepL	L'acteur, **bien que** beau et charmant et sélectionné parmi des centaines d'autres pour le prochain film de science-fiction, n'était pas une personne particulièrement agréable avec qui travailler.
S5h	Source	The hotel room that I managed to book at the last minute for a road trip, **while** it was clean and cheap, was not particularly attractive.
	Ref	La chambre d'hôtel que j'ai réussi à réserver à la dernière minute pour mon escapade routière, **bien que** propre et abordable, n'était pas particulièrement attrayante.
✗	Hybrid SMT	L' hôtel que j' ai réussi à réserver à la dernière minute pour un voyage de route, **alors qu'** elle était propre et bon marché, n' a pas été particulièrement attrayante.
✗*	CNN	La chambre d' hôtel que j' ai pu réserver à la dernière minute pour un voyage routier. L' hôtel est très bien situé, à proximité du centre ville.
✗	RNN	La chambre que j' ai réussi à réserver à la dernière minute pour un itinéraire, **alors qu'** elle était propre et bon marché, n' était pas particulièrement attrayante.
✗*	Attention	L' hôtel est très bien situé, très bien situé, à proximité de l' aéroport, de la gare et du centre ville.
✗	Google	La chambre d'hôtel que j'ai réussi à réserver à la dernière minute pour un road trip, **alors qu**'elle était propre et bon marché, n'était pas particulièrement attrayante.
✗	DeepL	La chambre d'hôtel que j'ai réussi à réserver à la dernière minute pour un voyage en voiture, **alors qu**'elle était propre et bon marché, n'était pas particulièrement attrayante.

"While" expressing an opposition

Possible solutions: *alors que*, *mais*, *quant à*, or *tandis que*

Short sentences		
"while" before the propositions		
S6a	Source	**While** Megan loves a good steak, her brother doesn't eat meat.
	Ref	**Tandis que** Megan adore un bon steak, son frère ne mange pas de viande.
✗	Hybrid SMT	**Bien que** Megan aime un bon steak, son frère doesn't manger de la viande.
✗	CNN	**Même si** Megan aime un bon steak, son frère ne mange pas La viande.
✗	RNN	**Bien que** Megan adore un bon steak, son frère ne mange pas de viande.
✓	Attention	**Tandis que** Megan aime un bon steak, son frère ne mange pas de viande.
✓	Google	**Alors que** Megan aime un bon steak, son frère ne mange pas de viande.
✗	DeepL	**Bien que** Megan aime un bon steak, son frère ne mange pas de viande.
S6b	Source	**While** the government is telling people to practice physical distancing, some still gather in parks.
	Ref	**Alors que** le gouvernement demande aux gens de pratiquer la distanciation physique, certains se rassemblent toujours dans les parcs.
✓	Hybrid SMT	**Alors que** le gouvernement est à dire aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.
✗	CNN	**Même si** le gouvernement dit aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.
✗	RNN	**Bien que** le gouvernement dise aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.
✗	Attention	**Bien que** le gouvernement dise aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.

✓	Google	**Alors que** le gouvernement dit aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.
✗	DeepL	**Bien que** le gouvernement dise aux gens de pratiquer la distanciation physique, certains se rassemblent encore dans les parcs.
"while" between the propositions		
S6c	Source	There was still soap on the shelves, **while** the toilet paper was out of stock.
	Ref	Il y avait encore du savon sur les étagères, **alors que** le papier toilette était en rupture de stock.
✓	Hybrid SMT	Il y avait encore du savon sur les tablettes, **tandis que** le papier de toilette était hors stock.
✓	CNN	Il y avait toujours du savon sur les étagères, **alors que** le papier de toilette était hors de stock.
✓	RNN	Il y avait encore du savon sur les étagères, **alors que** le papier de toilette était hors de stock.
✓	Attention	Il y avait encore du savon sur les étagères, **alors que** le papier de toilette était hors stock.
✓	Google	Il y avait encore du savon sur les étagères, **tandis que** le papier toilette était en rupture de stock.
✓	DeepL	Il y avait encore du savon sur les étagères, **alors que** le papier toilette était épuisé.
S6d	Source	The blue mushroom shrinks Mario, **while** the red one makes him grow.
	Ref	Le champignon bleu fait rétrécir Mario, **alors que** le rouge le fait grandir.
✓	Hybrid SMT	Le champignon bleu rétrécit Mario, **tandis que** le rouge lui permet de croître.
✓	CNN	Le champignon bleu diminue Mario, **alors que** le rouge le pousse.
✓	RNN	Le champignon bleu rétrécit Mario, **alors que** l' un rouge le fait croître.
✓	Attention	Le champignon bleu rétrécit Mario, **tandis que** celui rouge le fait grandir.
✓	Google	Le champignon bleu rétrécit Mario, **tandis que** le rouge le fait grandir.
✓	DeepL	Le champignon bleu rétrécit Mario, **tandis que** le rouge le fait grandir.

Long sentences		
"while" before the propositions		
S6e	Source	**While** Megan loves a good medium-rare steak with creamy peppercorn sauce, her younger brother who grew up with her doesn't eat meat at all.
	Ref	**Alors que** Megan adore un bon steak mi-saignant avec une sauce au poivre crémeuse, son petit frère, qui a grandi avec elle, ne mange pas du tout de viande.
✗	Hybrid SMT	**Bien que** Megan aime un bon steak mi-saignant avec la sauce crémeuse de poivre, son jeune frère qui a grandi avec elle doesn't manger de la viande.
✗	CNN	**Même si** Megan aime un bon steak à moyen terme avec une sauce au poivre crémeux, son frère cadet qui a grandi avec elle ne mange pas de viande.
✓	RNN	**Alors que** Megan adore un bon steak moyen-rare avec sauce poivrée poivrée, son jeune frère qui a grandi avec elle ne mange pas de viande du tout.
✓	Attention	**Tandis que** Megan aime un bon steak moyen-rere avec une sauce aux glands poivrés crémeuse, son plus jeune frère qui a grandi avec elle ne mange pas du tout de viande.
✓	Google	**Alors que** Megan aime un bon steak mi-saignant avec une sauce crémeuse au poivre, son jeune frère qui a grandi avec elle ne mange pas du tout de viande.
✓	DeepL	**Alors que** Megan aime un bon steak à point avec une sauce crémeuse au poivre, son jeune frère qui a grandi avec elle ne mange pas de viande du tout.
S6f	Source	**While** the government is telling people to practice physical distancing to reduce the spread of the virus, some still gather in parks and visit relatives.
	Ref	**Alors que** le gouvernement demande aux gens de pratiquer la distanciation physique pour réduire propagation du virus, certains se rassemblent toujours dans les parcs et visitent toujours leurs proches.
✓	Hybrid SMT	**Alors que** le gouvernement est à dire aux gens de pratiquer la distanciation physique afin de réduire la propagation du virus, certains se rassemblent encore dans les parcs et rendre visite à des parents.

✗	CNN	**Même si** le gouvernement dit aux gens qu' il prenne des mesures de distanciation physique pour réduire la propagation du virus Certains se rassemblent encore dans les parcs et visitent des parents.
✗	RNN	**Bien que** le gouvernement dise aux gens de pratiquer la distanciation physique pour réduire la propagation du virus, certains se rassemblent encore dans les parcs et visitent les parents.
✗	Attention	**Bien que** le gouvernement dise aux gens de pratiquer la distanciation physique pour réduire la propagation du virus, certains se rassemblent encore dans les parcs et visitent des parents.
✓	Google	**Alors que** le gouvernement dit aux gens de pratiquer la distanciation physique pour réduire la propagation du virus, certains se rassemblent encore dans des parcs et rendent visite à des proches.
✓	DeepL	**Alors que** le gouvernement dit aux gens de pratiquer la distanciation physique pour réduire la propagation du virus, certains se rassemblent encore dans les parcs et rendent visite à des parents.
"while" between the propositions		
S6g	Source	I lost faith in humanity when I saw that there was still hand soap on the shelves, **while** the toilet paper was out of stock.
	Ref	J'ai perdu foi en l'humanité quand j'ai vu qu'il y avait encore du savon à mains sur les étagères, **alors que** le papier toilette était en rupture de stock.
✓	Hybrid SMT	J' ai perdu la foi dans l' humanité quand j' ai vu qu' il y avait encore du savon à la main sur les tablettes, **tandis que** le papier de toilette était hors stock.
✗*	CNN	J' ai perdu confiance en l' humanité quand j' ai vu qu' il y avait encore du savon à main sur les étagères. Le papier de toilette n' était pas en stock.
✓	RNN	J' ai perdu foi en l' humanité quand j' ai vu qu' il y avait encore du savon à la main sur les étagères, **alors que** le papier de toilette était hors de stock.
✓	Attention	J' ai perdu confiance en l' humanité lorsque j' ai vu qu' il y avait encore du savon à la main sur les étagères, **alors que** le papier de toilette était hors stock.

✓	Google	J'ai perdu confiance en l'humanité quand j'ai vu qu'il y avait encore du savon pour les mains sur les étagères, **alors que** le papier toilette était en rupture de stock.
✓	DeepL	J'ai perdu foi en l'humanité quand j'ai vu qu'il y avait encore du savon pour les mains sur les étagères, **alors que** le papier toilette était épuisé.
S6h	Source	The blue mushroom shrinks Mario and allows him to walk on water, **while** the red one makes him bigger and more resistant to enemy attacks.
	Ref	Le champignon bleu fait rétrécir Mario et lui permet de marcher sur l'eau, **alors que** le rouge le fait grandir et le rend plus résistant aux attaques ennemies.
✓	Hybrid SMT	Le champignon bleu se rétrécit de Mario et lui permet de marcher sur l' eau, **tandis que** le rouge le rend de plus en plus résistants aux attaques ennemies.
✓	CNN	Le champignon bleu rétrécit Mario et lui permet de marcher sur l' eau, **tandis que** le rouge le fait Il est plus grand et plus résistant aux attaques ennemies.
✓	RNN	Le champignon bleu rétrécit Mario et lui permet de marcher sur l' eau, **alors que** le rouge le rend plus grand et plus résistant aux attaques ennemies.
✓	Attention	Le champignon bleu rétrécit Mario et lui permet de marcher sur l' eau, **tandis que** le champignon rouge le rend plus grand et plus résistant aux attaques ennemies.
✓	Google	Le champignon bleu rétrécit Mario et lui permet de marcher sur l'eau, **tandis que** le champignon rouge le rend plus gros et plus résistant aux attaques ennemies.
✓	DeepL	Le champignon bleu rétrécit Mario et lui permet de marcher sur l'eau, **tandis que** le rouge le rend plus grand et plus résistant aux attaques ennemies.

"When" expressing causality

Possible solutions: *par suite de*, *pour*, or *parce que*, or through a sentence structure that conveys causality

Short sentences		
"when" before the propositions		
S7a	Source	**When** the freezers stopped working, the store had to throw away all its ice cream.
	Ref	**À cause de** la panne des congélateurs, le magasin a dû jeter toute sa crème glacée.
✗	Hybrid SMT	**Lorsque** les congélateurs ont cessé de fonctionner, le magasin avait de jeter toutes ses glaces.
✗	CNN	**Lorsque** les congélateurs ont cessé de travailler, le magasin a dû jeter toute sa crème glacée.
✗	RNN	**Lorsque** les congélateurs ont cessé de fonctionner, le magasin a dû jeter toute sa crème glacée.
✗	Attention	**Lorsque** les congélateurs ont cessé de travailler, le magasin a dû jeter toute sa crème glacée.
✗	Google	**Lorsque** les congélateurs ont cessé de fonctionner, le magasin a dû jeter toute sa glace.
✗	DeepL	**Lorsque** les congélateurs ont cessé de fonctionner, le magasin a dû jeter toutes ses glaces.
S7b	Source	**When** the oil spilled, dozens of otter dens were completely destroyed.
	Ref	**À cause du** déversement de pétrole, des dizaines de tanières de loutres ont été complètement détruites.
✗	Hybrid SMT	**Lorsque** le pétrole déversé, des dizaines de tanières de loutres ont été complètement détruits.
✗	CNN	**Lorsque** l' huile s' est déversée, des dizaines de terriers de loutres ont été complètement détruits.
✗	RNN	**Lorsque** l' huile s' est déversée, des douzaines de terriers de loutres ont été complètement détruits.
✗	Attention	**Lorsque** l' huile s' est déversée, des dizaines de tanières de loutres ont été complètement détruites.
✗	Google	**Lorsque** le pétrole s'est déversé, des dizaines de tanières de loutres ont été complètement détruites.
✗	DeepL	**Lorsque** le pétrole s'est répandu, des dizaines de tanières de loutre ont été complètement détruites.

"when" between the propositions		
S7c	Source	She was diagnosed with epilepsy **when** she collapsed and had a seizure.
	Ref	Elle a reçu un diagnostic d'épilepsie **parce qu'**elle s'est évanouie et a eu des convulsions.
✗	Hybrid SMT	Elle a reçu un diagnostic d' épilepsie **lorsqu'** elle s' est effondré et a eu une crise.
✗	CNN	On lui a diagnostiqué l' épilepsie **lorsqu'** elle s' est effondrée et a subi une crise.
✗	RNN	On lui a diagnostiqué une épilepsie **lorsqu'** elle s' est effondrée et qu' elle a été saisie.
✗	Attention	On lui a diagnostiqué l' épilepsie **lorsqu'** elle s' est effondrée et qu' elle a subi une crise.
✗	Google	Elle a reçu un diagnostic d'épilepsie **lorsqu'**elle s'est effondrée et a eu une crise.
✗	DeepL	On lui a diagnostiqué de l'épilepsie **lorsqu'**elle s'est effondrée et a eu une crise.
S7d	Source	The neighbours stepped in to donate furniture and clothing **when** the apartment building caught fire.
	Ref	Les voisins ont fait des dons de meubles et de vêtements **par suite de** l'incendie du bloc appartment.
✗	Hybrid SMT	Les voisins sont intervenus à donner des meubles et des vêtements **lorsque** l' immeuble a pris feu.
✗	CNN	Les voisins s' efforcent de donner du mobilier et de l' habillement **lorsque** l' immeuble d' habitation a pris feu.
✗	RNN	Les voisins se sont empressés de donner des meubles et des vêtements **lorsque** l' immeuble a pris feu.
✗	Attention	Les voisins ont fait don de meubles et de vêtements **lorsque** l' immeuble d' appartements a pris feu.
✗	Google	Les voisins sont intervenus pour donner des meubles et des vêtements **lorsque** l'immeuble a pris feu.
✗	DeepL	Les voisins sont intervenus pour donner des meubles et des vêtements **lorsque** l'immeuble a pris feu.

Long sentences		
"when" before the propositions		
S7e	Source	**When** the six brand-new freezers in aisle 3 suddenly stopped working all at once, the grocery store had to throw away all its ice cream.
	Ref	**À cause de** la panne soudaine et simultanée des six nouveaux congélateurs dans l'allée 3, l'épicerie a dû jeter toute sa crème glacée.
✗	Hybrid SMT	**Lorsque** les six congélateurs flambant dans le couloir 3 a soudainement cessé de travailler tout à la fois, l' épicerie a dû jeter toutes ses glaces.
✗	CNN	**Lorsque** les six nouveaux congélateurs dans l' allée 3 ont cessé de travailler tout à la fois L' épicerie a dû jeter toute sa crème glacée.
✗	RNN	**Lorsque** les six nouveaux congélateurs dans l' allée 3 ont soudainement cessé de travailler en même temps, l' épicerie a dû jeter toute sa crème glacée.
✗	Attention	**Lorsque** les six nouveaux congélateurs de l' allée 3 ont soudainement cessé de travailler tout à la fois, l' épicerie a dû jeter toute sa crème glacée.
✗	Google	**Lorsque** les six congélateurs flambant neufs de l'allée 3 ont soudainement cessé de fonctionner, l'épicerie a dû jeter toute sa glace.
✗	DeepL	**Lorsque** les six congélateurs flambant neufs de l'allée 3 ont soudainement cessé de fonctionner d'un seul coup, l'épicerie a dû jeter toutes ses glaces.
S7f	Source	**When** the oil arriving from Alaska in the United States accidentally spilled on the coasts of the Pacific Ocean, dozens of otter dens were completely destroyed.
	Ref	**À cause du** déversement accidentel de pétrole, des dizaines de loutres se sont retrouvées sans abri parce que le pétrole provenant des États-Unis s'est accidentellement déversé sur les côtes de l'est du Pacifique Nord.
✗	Hybrid SMT	**Lorsque** le pétrole en provenance de l' Alaska aux États-Unis déversé accidentellement sur les côtes de l' océan Pacifique, des dizaines de tanières de loutres ont été complètement détruits.
✗	CNN	**Lorsque** l' huile arrivant d' Alaska aux États-Unis accidentellement déversée sur les côtes Océan Pacifique, des dizaines de terriers de loutres ont été complètement détruits.

✗	RNN	**Lorsque** l' huile provenant de l' Alaska aux États-Unis s' est déversée accidentellement sur les côtes de l' océan Pacifique, des douzaines de loutres ont été complètement détruites.
✗	Attention	**Lorsque** le pétrole arrivant d' Alaska aux États-Unis a accidentellement déversé sur les côtes de l' océan Pacifique, des dizaines de tanières de loutres ont été complètement détruites.
✗	Google	**Lorsque** le pétrole arrivant d'Alaska aux États-Unis s'est accidentellement déversé sur les côtes de l'océan Pacifique, des dizaines de tanières de loutres ont été complètement détruites.
✗	DeepL	**Lorsque** le pétrole arrivant d'Alaska aux États-Unis s'est accidentellement déversé sur les côtes de l'océan Pacifique, des dizaines de tanières de loutres ont été complètement détruites.
"when" between the propositions		
S7g	Source	She was diagnosed with epilepsy, leaving her unable to participate in high-risk sports such as scuba and skydiving, **when** she collapsed and had a seizure at the age of 8.
	Ref	Elle a reçu un diagnostic d'épilepsie, ce qui l'empêche de pratiquer des sports à haut risque tels que la plongée et le parachutisme, **à la suite de** s'être évanouie et d'avoir eu des convulsions à l'âge de 8 ans.
✗	Hybrid SMT	Elle a reçu un diagnostic d' épilepsie, laissant son incapacité à participer à des sports à haut risque comme la plongée et le parachutisme, **lorsqu'** elle s' est effondré et a eu une crise à l' âge de 8 ans.
✗	CNN	Elle a été diagnostiquée d' épilepsie, laissant elle incapable de participer à des sports à haut risque comme la plongée sous-marine et la parachute, **lorsqu ' ** elle s ' est effondrée et a été saisie au L' âge de 8 ans.
✗	RNN	Elle a reçu un diagnostic d' épilepsie, laissant elle incapable de participer à des sports à haut risque, comme la plongée sous-marine et le parachutisme, **lorsqu'** elle s' est effondrée.
✗	Attention	On lui a diagnostiqué l' épilepsie, laissant son incapacité à participer à des sports à haut risque comme la plongée sous-marine et la plongée sous-marine, **lorsqu'** elle s' est effondue.

✗	Google	Elle a reçu un diagnostic d'épilepsie, la laissant incapable de participer à des sports à haut risque comme la plongée sous-marine et le parachutisme, **lorsqu'**elle s'est effondrée et a eu une crise à l'âge de 8 ans.
✗	DeepL	On lui a diagnostiqué une épilepsie qui l'a empêchée de pratiquer des sports à haut risque comme la plongée sous-marine et le parachutisme, **lorsqu'**elle s'est effondrée et a eu une crise à l'âge de 8 ans.
S7h	Source	The neighbours who just moved in next-door last month stepped in to donate food, furniture and clothing **when** the apartment building caught fire in the middle of the night.
	Ref	Les voisins qui viennent d'emménager à côté le mois dernier ont fait des dons de nourriture, de meubles et de vêtements **par suite de** l'incendie du bloc appartment qui a eu lieu au milieu de la nuit.
✗	Hybrid SMT	Les voisins qui ont tout juste emménagé le mois dernier voisin est intervenu pour donner de la nourriture, des meubles et des vêtements **lorsque** l' immeuble a pris feu au milieu de la nuit.
✗	CNN	Les voisins qui viennent de déménager à la porte le mois dernier s' étaient déplacés pour donner des aliments, des meubles et des vêtements **lorsque** l' immeuble d' appartement a pris feu au milieu de la nuit.
✗	RNN	Les voisins qui venaient tout juste de déménager le mois dernier ont fait un don de nourriture, de mobilier et de vêtements **lorsque** l' immeuble a pris feu au milieu de la nuit.
✗	Attention	Les voisins qui viennent de s' installer à la porte du mois dernier ont fait don de nourriture, de meubles et de vêtements **lorsque** l' immeuble d' appartements a pris feu au milieu de la nuit.
✗	Google	Les voisins qui viennent d'emménager le mois dernier sont intervenus pour donner de la nourriture, des meubles et des vêtements **lorsque** l'immeuble a pris feu au milieu de la nuit.
✗	DeepL	Les voisins qui viennent d'emménager à côté le mois dernier sont intervenus pour donner de la nourriture, des meubles et des vêtements **lorsque** l'immeuble a pris feu au milieu de la nuit.

"When" expressing continuity
Possible solutions: *et* or *puis*

Short sentences		
"when" before the propositions		
S8a	Source	The helicopter slowly landed in the field, **when** the rescuers jumped off.
	Ref	L'hélicoptère a atteri lentement dans le champ, **puis** les sauveteurs ont sauté de l'hélicoptère.
✗	Hybrid SMT	L' hélicoptère a atterri lentement dans le domaine, **lorsque** les sauveteurs ont sauté.
✗	CNN	L' hélicoptère débarquait lentement sur le terrain, **lorsque** les sauveteurs ont sauté.
✗	RNN	L ' hélicoptère s ' est débarrassé lentement dans le champ, **lorsque** les sauveteurs ont sauté.
✗	Attention	L' hélicoptère s' est lentement débarqué sur le terrain, **lorsque** les sauveteurs ont sauté.
✗	Google	L'hélicoptère a atterri lentement sur le terrain, **lorsque** les sauveteurs ont sauté.
✗	DeepL	L'hélicoptère s'est lentement posé sur le terrain, **lorsque** les sauveteurs ont sauté.
S8b	Source	The car will be fixed by 5pm, **when** I will come to pick it up.
	Ref	La voiture sera réparée d'ici 17 h **et** je viendrai la chercher.
✗	Hybrid SMT	La voiture sera fixée par 5pm, **lorsque** je viendrai à le ramasser.
✗	CNN	La voiture sera fixée à 17h, **lorsque** je vais venir le ramasser.
✗	RNN	La voiture sera fixée à 17h, **alors que** je viendrai le chercher.
✗	Attention	La voiture sera fixée d' ici 17h, **quand** je vais venir la ramasser.
✗	Google	La voiture sera réparée à 17 heures, **lorsque** je viendrai la chercher.
✗	DeepL	La voiture sera réparée à 17 heures, **quand** je viendrai la chercher.

"when" between the propositions		
S8c	Source	I drove until it was dark outside, **when** my husband took the wheel.
	Ref	J'ai conduit jusqu'à ce qu'il fasse noir dehors, **puis** mon mari a pris le volant.
✗	Hybrid SMT	J' ai conduit jusqu' à ce qu' elle était sombre à l' extérieur, **lorsque** mon mari a pris la roue.
✗	CNN	J' ai conduit jusqu' à ce qu' il n' ait pas été foncé, **lorsque** mon mari a pris la roue.
✗	RNN	J' ai conduit jusqu' à ce qu' il soit sombre à l' extérieur, **lorsque** mon mari a pris la roue.
✗	Attention	J' ai conduit jusqu' à ce qu' il soit sombre à l' extérieur, **lorsque** mon mari a pris la roue.
✗	Google	J'ai conduit jusqu'à ce qu'il fasse nuit dehors, **quand** mon mari a pris le volant.
✗	DeepL	J'ai conduit jusqu'à ce qu'il fasse nuit dehors, **quand** mon mari a pris le volant.
S8d	Source	They swam until their fingertips were wrinkly, **when** they got out of the pool.
	Ref	Ils ont nagé jusqu'à ce que le bout de leurs doigts soit fripé, **puis** sont sortis de la piscine.
✗	Hybrid SMT	Ils nagent jusqu' à ce que leurs doigts ont été wrinkly, **alors qu'** ils sortaient de la piscine.
✗	CNN	Ils nageaient jusqu' à ce que leurs doigts soient ridicules, **alors qu'** ils s' en sortent. La piscine est située à proximité de la piscine.
✗	RNN	Ils nageaient jusqu' à ce que leurs bouts de doigts soient ridicules, **lorsqu'** ils sortaient de la piscine.
✗	Attention	Ils s' écoulent jusqu' à ce que leurs doigts soient troublés, **lorsqu'** ils sortent de la piscine.
✗	Google	Ils ont nagé jusqu'à ce que leurs doigts soient froissés, **lorsqu'**ils sont sortis de la piscine.
✗	DeepL	Ils ont nagé jusqu'à ce que le bout de leurs doigts soit plissé, **quand** ils sont sortis de la piscine.

Long sentences		
"when" before the propositions		
S8e	Source	The helicopter slowly landed in the field below the big rock wall, **when** the three rescuers jumped off and went looking for the injured climbers.
	Ref	L'hélicoptère a atteri lentement dans le champ au bas de la grande paroi rocheuse, **puis** les trois sauveteurs ont sauté de l'hélicoptère et sont allés chercher les alpinistes blessés.
✗	Hybrid SMT	L' hélicoptère a atterri lentement dans le champ ci-dessous le mur de Big rock, **lorsque** les trois sauveteurs ont sauté et partit à la recherche des grimpeurs blessés.
✗	CNN	L' hélicoptère débarquait lentement dans le champ sous le gros mur rocheux, **lorsque** les trois sauveteurs s ' est écrasé et s ' est rendu à la recherche des alpinistes blessés.
✗	RNN	L ' hélicoptère s ' est débarrassé lentement dans le champ sous le gros mur de roche, **lorsque** les trois sauveteurs ont sauté et ont cherché les grimpeurs blessés.
✗	Attention	L' hélicoptère s' est atterri lentement dans le champ sous le grand mur rocheux, **lorsque** les trois sauveteurs ont sauté et ont cherché les alpinistes blessés.
✗	Google	L'hélicoptère a atterri lentement dans le champ sous la grande paroi rocheuse, **lorsque** les trois sauveteurs ont sauté et sont allés à la recherche des grimpeurs blessés.
✗	DeepL	L'hélicoptère a lentement atterri dans le champ en dessous de la grande paroi rocheuse, **quand** les trois sauveteurs ont sauté et sont partis à la recherche des alpinistes blessés.
S8f	Source	The mechanic who was recommended to me by a colleague will have the car fixed by 5pm, **when** I will come to pick it up.
	Ref	Le mécanicien que m'a recommandé un collègue aura réparé la voiture d'ici 17 h et je pourrai **alors** venir la chercher.
✗	Hybrid SMT	Le mécanicien qui m' a été recommandé par un collègue aura la voiture fixé par 5pm, **lorsque** je viendrai à le ramasser.
✗	CNN	Le mécanicien qui m' a recommandé par un collègue aura la voiture fixée à 17 h, **quand** J' arriverai à la ramasser.
✗	RNN	Le mécanicien qui m' a été recommandé par un collègue aura la voiture fixée à 17 heures, **alors que** je viens le chercher.

✗	Attention	Le mécanicien qui m' a été recommandé par un collègue aura la voiture fixée d' ici 17h, **quand** je vais la ramasser.
✗	Google	Le mécanicien qui m'a été recommandé par un collègue fera réparer la voiture avant 17h, **lorsque** je viendrai la chercher.
✗	DeepL	Le mécanicien qui m'a été recommandé par un collègue fera réparer la voiture avant 17 heures, **quand** je viendrai la chercher.
"when" between the propositions		
S8g	Source	I drove my aunt's old car with two of my friends in the backseat and my husband next to me until it was dark outside, **when** my husband took the wheel.
	Ref	J'ai conduit la vieille voiture de ma tante avec deux amis assis en arrière et mon mari à côté de moi jusqu'à ce qu'il fasse noir dehors, **puis** mon mari a pris le volant.
✗	Hybrid SMT	J' ai conduit ma tante 's vieille voiture avec deux de mes amis sur le siège arrière et mon mari à côté de moi, jusqu' à ce qu' elle était sombre à l' extérieur, **lorsque** mon mari a pris la roue.
✗	CNN	J' ai conduit la vieille voiture de ma tante avec deux de mes amis dans l' arrière-place Mon mari m' a côté de moi jusqu' à ce qu' il n' ait été foncé à l' extérieur, **lorsque** mon mari a pris la roue.
✗	RNN	J' ai conduit la vieille voiture de ma tante avec deux de mes amis dans le siège arrière et mon mari à côté de moi jusqu' à ce qu' il soit foncé, **lorsque** mon mari a pris la roue.
✗	Attention	J' ai conduit la vieille voiture de ma tante avec deux de mes amis dans le siège arrière et mon mari près de moi jusqu' à ce qu' elle soit sombre à l' extérieur, **lorsque** mon mari a pris la roue.
✗	Google	J'ai conduit la vieille voiture de ma tante avec deux de mes amis sur la banquette arrière et mon mari à côté de moi jusqu'à ce qu'il fasse noir dehors, **lorsque** mon mari a pris le volant.
✗	DeepL	J'ai conduit la vieille voiture de ma tante avec deux de mes amis sur le siège arrière et mon mari à côté de moi jusqu'à ce qu'il fasse nuit dehors, **quand** mon mari a pris le volant.

S8h	Source	They swam all afternoon at Katie's new house until their fingertips were wrinkly and they were all shivering, **when** they got out of the pool.
	Ref	Ils ont nagé toute l'après-midi à la nouvelle maison de Katie jusqu'à ce que le bout de leurs doigts soit fripé et qu'ils tremblaient tous, **puis** sont sortis de la piscine.
✗	Hybrid SMT	Ils nageaient tous les après-midi à Katie 's nouvelle maison jusqu' à ce que leurs doigts ont été wrinkly et ils étaient tous le frisson, **alors qu'** ils sortaient de la piscine.
✗	CNN	Ils nageaient tous l' après-midi à la nouvelle maison de Katie jusqu' à ce qu' il y ait des doigts Ils étaient ridicules et ils étaient tous frissons, **quand** ils s ' étaient retirés de la s' il s' agit d' un groupe d' experts.
✗	RNN	Ils nageaient tout l' après-midi à la nouvelle maison de Katie jusqu' à ce que leurs bouts de doigts soient ridicules et qu' ils soient tous frissonnés, **lorsqu'** ils sortent de la piscine.
✗	Attention	Ils s' éteignent tout l' après-midi à la nouvelle maison de Katie jusqu' à ce que leurs doigts soient troublés et qu' ils soient tous frissonnants **quand** ils sortent.
✗	Google	Ils ont nagé toute l'après-midi dans la nouvelle maison de Katie jusqu'à ce que leurs doigts soient froissés et qu'ils tremblent tous, **quand** ils sont sortis de la piscine.
✗	DeepL	Ils ont nagé tout l'après-midi dans la nouvelle maison de Katie jusqu'à ce que le bout de leurs doigts soit ridé et qu'ils frissonnent tous, **lorsqu'**ils sont sortis de la piscine.

"When" meaning "in spite of the fact that"

Possible solutions: *alors que*, *malgré le fait que*, *en dépit du fait que*, etc.

Short sentences		
S9a	Source	I was angry that he left **when** he said he would stay to help.
	Ref	J'étais fâché qu'il soit parti **malgré le fait qu'**il ait dit qu'il resterait aider.
✗	Hybrid SMT	J' étais en colère qu' il a quitté **lorsqu'** il a dit qu' il allait rester pour aider.
✗	CNN	J' étais en colère qu' il partait **lorsqu'** il a dit qu' il allait continuer d' aider.
✗	RNN	J' étais en colère qu' il est parti **lorsqu'** il a dit qu' il resterait à l' aise.

✗	Attention	J' étais en colère **lorsqu'** il a dit qu' il restait à aider.
✗	Google	J'étais en colère qu'il soit parti **quand** il a dit qu'il resterait pour aider.
✗	DeepL	J'étais furieux qu'il soit parti **quand** il a dit qu'il resterait pour aider.
S9b	Source	They couldn't believe he made that decision **when** he knew all the risks it involved.
	Ref	Ils n'arrivaient pas à croire qu'il avait pris cette décision **alors qu'**il connaissait tous les risques qui en découlaient.
✓	Hybrid SMT	Ils croient couldn't il a pris cette décision **alors qu'** il savait tous les risques qu' elle implique.
✗	CNN	Ils ne pouvaient pas croire qu' il a pris cette décision **lorsqu'** il connaissait tous les risques qu' il comporte.
✗	RNN	Ils ne pouvaient pas croire qu' il avait pris cette décision **lorsqu'** il connaissait tous les risques qu' il comportait.
✗	Attention	Ils ne pouvaient croire qu' il avait pris cette décision **lorsqu'** il connaissait tous les risques qu' elle comporte.
✓	Google	Ils ne pouvaient pas croire qu'il avait pris cette décision **alors qu'**il connaissait tous les risques que cela impliquait.
✓	DeepL	Ils ne pouvaient pas croire qu'il avait pris cette décision **alors qu'**il connaissait tous les risques que cela impliquait.
S9c	Source	I was upset she still dated him **when** I warned her he was a cheater.
	Ref	J'étais contrarié qu'elle sorte encore avec lui **alors que** je l'avais avertie qu'il était un trompeur.
✗	Hybrid SMT	J' ai été bouleversée, elle lui fait encore **quand** j' ai averti qu' il était un tricheur.
✗*	CNN	J' ai été bouleversée qu' elle l' avait toujours avertie qu' il était un tricheur.
✗	RNN	J' étais bouleversée qu' elle l' ait encore datée **lorsque** j' ai averti qu' il était un tricheur.
✗	Attention	J' étais bouleversée qu' elle le daignait **quand** je l' ai avertie qu' il était tricheur.
✗	Google	J'étais bouleversée qu'elle soit encore sortie avec lui **quand** je l'ai prévenue qu'il était un tricheur.
✗	DeepL	J'étais contrarié qu'elle soit encore sortie avec lui **quand** je l'ai averti qu'il était un tricheur.

S9d	Source	It made me sick to see him defend her **when** she's been treating him like garbage.
	Ref	Ça m'a rendu malade de le voir la défendre **alors qu'**elle le traitait comme un déchet.
✗	Hybrid SMT	Cela m' a fait malade de le voir défendre son 's **quand** elle a été le traitant comme des ordures.
✗	CNN	Cela m' a rendu malade pour le voir se défendre **lorsqu'** elle le traitait comme des ordures.
✗	RNN	Il m' a rendu malade de le voir la défendre **lorsqu'** elle l' a traitée comme des ordures.
✗	Attention	Il m' a rendu malade de le voir défendre **lorsqu'** elle le traitait comme des ordures.
✓	Google	Cela m'a rendu malade de le voir la défendre **alors qu'**elle le traitait comme une poubelle.
✓	DeepL	Ça me rendait malade de le voir la défendre **alors qu'**elle le traitait comme un déchet.
Long sentences		
S9e	Source	I was angry that he left without telling anyone and with all of the tools he knew we needed, **when** he said he would stay to help.
	Ref	J'étais fâché qu'il soit parti sans le dire à personne et avec tous les outils dont il savait que nous avions besoin, **alors qu'**il avait dit qu'il resterait aider.
✗	Hybrid SMT	J' étais en colère qu' il a laissé sans le dire à quiconque et avec tous les outils qu' il savait que nous avions besoin, **quand** il a dit qu' il allait rester pour aider.
✗	CNN	J' étais en colère qu' il n' a pas dit à qui que ce soit et avec tous les outils qu' il savait, **lorsqu'** il a dit qu' il resterait à l' aide.
✗	RNN	J' étais en colère qu' il s' est laissé sans dire à qui que ce soit et avec tous les outils dont il avait besoin, **quand** il a dit qu' il resterait pour aider.
✗	Attention	J' étais en colère qu' il partait sans dire à qui que ce soit et avec tous les outils qu' il savait que nous avions besoin, **lorsqu'** il a dit qu' il resterait à aider.
✗	Google	J'étais en colère qu'il soit parti sans le dire à personne et avec tous les outils dont il savait que nous avions besoin, **quand** il a dit qu'il resterait pour aider.

✓	DeepL	J'étais en colère qu'il soit parti sans le dire à personne et avec tous les outils dont il savait que nous avions besoin, **alors qu'**il avait dit qu'il resterait pour nous aider.
S9f	Source	They couldn't believe he made that decision without consulting anyone from management or from his working group, **when** he knew all the risks it involved.
	Ref	Ils n'arrivaient pas à croire qu'il avait pris cette décision sans consulter qui que ce soit de la direction ou de son équipe de travail, **alors qu'**il connaissait tous les risques qui en découlaient.
✓	Hybrid SMT	Ils croient couldn't il a pris cette décision sans consulter quiconque de gestion ou de son groupe de travail, **alors qu'** il savait tous les risques qu' elle implique.
✗	CNN	Ils ne pouvaient pas croire qu' il a pris cette décision sans consulter quelqu' un d' une direction ou de son groupe de travail, **lorsqu'** il connaissait tous les risques qu' il comporte.
✗	RNN	Ils ne pouvaient croire qu' il avait pris cette décision sans consulter quelqu' un de la direction ou de son groupe de travail, **lorsqu'** il connaissait tous les risques qu' il implique.
✓	Attention	Ils ne pouvaient croire qu' il a pris cette décision sans consulter quiconque de la direction ou de son groupe de travail, **alors qu'** il connaissait tous les risques qu' elle comporte.
✓	Google	Ils ne pouvaient pas croire qu'il avait pris cette décision sans consulter quiconque de la direction ou de son groupe de travail, **alors qu'**il connaissait tous les risques que cela impliquait.
✓	DeepL	Ils ne pouvaient pas croire qu'il avait pris cette décision sans consulter personne de la direction ou de son groupe de travail, **alors qu'**il connaissait tous les risques que cela impliquait.
S9g	Source	I was upset and our whole friend group was frustrated that she still trusted and dated him, **when** I warned her that he was a cheater
	Ref	J'étais contrarié et notre groupe d'amis entier était frustré qu'elle lui fasse encore confiance et sorte encore avec lui **alors que** je l'avais avertie qu'il était un trompeur.

✗	Hybrid SMT	J' ai été bouleversée et notre groupe de tout ami était mécontent qu' elle lui a toujours fait confiance et datée, **quand** j' ai avertie qu' il était un tricheur
✗	CNN	J' ai été bouleversé et notre groupe d' amis était frustré qu' elle avait toujours confiance et qu' elle l' avait datée, **lorsque** je l' ai averti qu' il était un tricheur
✗	RNN	J ' étais bouleversée et tout notre groupe d ' amis était frustré qu ' elle l ' ait toujours fait confiance et datée, **lorsque** j ' ai averti qu ' il était un tricheur
✗	Attention	J' étais bouleversée et tout notre groupe d' amis était frustré qu' elle ait toujours confiance et daté, **quand** je l' ai averti qu' il était un tricheur.
✗	Google	J'étais bouleversé et tout notre groupe d'amis était frustré qu'elle continue de lui faire confiance et de sortir avec lui, **quand** je l'ai avertie qu'il était un tricheur
✗	DeepL	J'étais contrariée et tout notre groupe d'amis était frustré qu'elle lui fasse encore confiance et sorte avec lui, **lorsque** je l'ai avertie qu'il était un tricheur
S9h	Source	It made me sick to my stomach to see him defend her in front of all of our friends **when** she's been treating him like garbage.
	Ref	Ça m'a rendu complètement malade de le voir la défendre devant tous nos amis **alors qu**'elle le traitait comme un déchet.
✗	Hybrid SMT	Cela m' a fait malade à mon estomac de le voir défendre devant tous nos amis **lorsqu'** elle a été 's le traitant comme des ordures.
✗	CNN	Elle m' a rendu malade à l' estomac pour le voir se défendre devant tous nos amis **lorsqu ' ** elle « s' occupe de lui comme des ordures.
✗	RNN	Il m' a rendu malade à mon estomac pour le voir se défendre devant tous nos amis **lorsqu'** elle l' a traitée comme des ordures.
✗	Attention	Il m' a rendu malade à l' estomac pour le voir défendre devant tous nos amis **lorsqu'** elle le traitait comme des déchets.
✓	Google	Cela m'a fait mal au ventre de le voir la défendre devant tous nos amis **alors qu'**elle le traitait comme une poubelle.
✓	DeepL	Ça me donnait mal au ventre de le voir la défendre devant tous nos amis **alors qu'**elle le traite comme un déchet.

"With" expressing causality

Possible solutions: *à cause de*, *en raison de*, *car*, *parce que*, *comme*, etc., or using a gerund

Short sentences		
"with" at the start of the sentence		
S10a	Source	**With** the neighbour's dog constantly barking, I couldn't get any sleep.
	Ref	**Comme** le chien du voisin aboyait constamment, je n'ai pas pu dormir du tout.
✗	Hybrid SMT	**Avec** le chien qui aboyait constamment 's voisin, je couldn't obtenez tout le sommeil.
✗	CNN	**Avec** le chien du voisin en permanence, je n' ai pas pu obtenir Le sommeil est laissé à l ' eau.
✗	RNN	**Avec** le chien du voisin, je n' arrivais pas à dormir.
✗	Attention	**Avec** le chien du voisin, je n' ai pas pu dormir sans cesse.
✗	Google	**Avec** le chien du voisin qui aboyait constamment, je ne pouvais pas dormir.
✗	DeepL	**Avec** le chien du voisin qui aboie constamment, je n'arrivais pas à dormir.
S10b	Source	**With** everyone asked to stay home, the price of gas has drastically decreased.
	Ref	**Comme** on a demandé à tout le monde de rester à la maison, le coût du pétrole a considérablement diminué.
✗*	Hybrid SMT	Tout le monde a demandé de rester à la maison, le prix du gaz a considérablement diminué.
✗	CNN	**Avec** tout le monde a demandé de rester à la maison, le prix du gaz a considérablement diminué.
✓	RNN	**Comme** tout le monde a demandé à rester à la maison, le prix du gaz a considérablement diminué.
✓	Attention	**Comme** chacun a demandé de rester chez lui, le prix du gaz a considérablement diminué.
✓	Google	Tout le monde **étant** invité à rester à la maison, le prix de l'essence a considérablement diminué.
✓	DeepL	**Comme** tout le monde est prié de rester à la maison, le prix de l'essence a considérablement diminué.

"with" mid-sentence		
S10c	Source	I can't finish any of my work **with** people pestering me all the time.
	Ref	Je ne peux terminer aucun de mes travaux **quand** on me harcèle constamment.
✗	Hybrid SMT	J' ai fini can't tout de mon travail **avec** les gens me pestering tout le temps.
✗	CNN	Je n' arrive pas à terminer mon travail **avec** des gens qui m' nuisent tout le temps.
✗	RNN	Je ne peux pas terminer un de mes travaux **avec** des gens qui m' ébranlent tout le temps.
✗	Attention	Je ne peux terminer aucun de mes travaux **avec** des gens qui me ravagent tout le temps.
✗	Google	Je ne peux terminer aucun de mes travaux **avec** des gens qui me harcèlent tout le temps.
✗	DeepL	Je ne peux pas finir mon travail **avec** des gens qui me harcèlent tout le temps.
S10d	Source	I'm thinking of riding my bicycle, **with** the price of gas on the rise.
	Ref	Je pense faire du vélo, le prix de l'essence **étant** à la hausse.
✗	Hybrid SMT	La pensée I'm d' équitation mon vélo, **avec** le prix du gaz à la hausse.
✗	CNN	Je pense à mon vélo, **avec** le prix du gaz à la hausse.
✗	RNN	Je pense à faire ma bicyclette, **avec** le prix du gaz à la hausse.
✗	Attention	Je pense à ma bicyclette, **avec** le prix du gaz en hausse.
✗	Google	Je pense à faire du vélo, **avec** le prix de l'essence à la hausse.
✗	DeepL	Je pense à faire du vélo, **avec** le prix de l'essence qui augmente.
Long sentences		
"with" at the start of the sentence		
S10e	Source	**With** the neighbour's dog—a four-year-old Australian Shepherd that was rescued from a local shelter two weeks ago—constantly barking, I couldn't get any sleep.
	Ref	**Comme** le chien du voisin, un berger australien de quatre ans qui a été adopté d'un refuge local voilà deux semaines, aboyait constamment, je n'ai pas pu dormir du tout.

✗	Hybrid SMT	**Avec** le chien 's voisin - un berger australien de quatre ans qui a été sauvée de refuge local il y a deux semaines - aboient constamment, je couldn't obtenez tout le sommeil.
✗	CNN	**Avec** le chien du voisin, un berger australien de quatre ans qui était sauvé d' un abri local il y a deux semaines, sans cesse, je ne pouvais pas n' importe quel sommeil.
✗	RNN	**Avec** le chien du voisin - un berger australien de quatre ans qui a été sauvé d' un abri local il y a deux semaines - sans constance, je ne pouvais pas dormir.
✗	Attention	**Avec** le chien du voisin, un berger australien de quatre ans qui a été sauvé d' un refuge local il y a deux semaines, je n' ai jamais pu dormir.
✗	Google	**Avec** le chien du voisin - un berger australien de quatre ans qui a été sauvé d'un refuge local il y a deux semaines - qui aboyait constamment, je ne pouvais pas dormir.
✗	DeepL	**Avec** le chien du voisin - un berger australien de quatre ans qui a été sauvé d'un refuge local il y a deux semaines - qui aboie sans cesse, je n'arrive pas à dormir.
S10f	Source	**With** everyone asked to stay home in hopes of reducing the number of cases of COVID-19 in Canada, the cost of gas has drastically decreased.
	Ref	**Comme** on a demandé à tout le monde de rester à la maison dans l'espoir de réduire le nombre de cas de COVID-19 au Canada, le coût du pétrole a considérablement diminué.
✗*	Hybrid SMT	Tout le monde a demandé de rester à la maison dans l' espoir de réduire le nombre de cas de COVID-19 au Canada, le coût de l' essence a considérablement diminué.
✗	CNN	**Avec** toutes les personnes demandées à rester à la maison dans l ' espoir de réduire le nombre de cas de COVID-19 en Le Canada, le coût du gaz a considérablement diminué.
✓	RNN	**Comme** tout le monde a demandé à rester à la maison dans l' espoir de réduire le nombre de cas de COVID-19 au Canada, le coût du gaz a diminué considérablement.
✓	Attention	**Comme** chacun a demandé de rester chez lui dans l' espoir de réduire le nombre de cas de COVID-19 au Canada, le coût du gaz a considérablement diminué.
✓	Google	Tout le monde **étant** invité à rester à la maison dans l'espoir de réduire le nombre de cas de COVID-19 au Canada, le coût de l'essence a considérablement diminué.

✓	DeepL	**Comme** tout le monde a été invité à rester à la maison dans l'espoir de réduire le nombre de cas de COVID-19 au Canada, le prix de l'essence a considérablement diminué.
"with" mid-sentence		
S10g	Source	I can't finish any of the work assigned to me by my new manager who started last week, **with** people pestering me all the time.
	Ref	Je ne peux terminer aucun des travaux que m'a confiés mon nouveau gestionnaire qui a commencé la semaine dernière, **quand** on me harcèle constamment.
✗	Hybrid SMT	J' ai fini can't aucune des tâches qui me sont confiées par mon nouveau gestionnaire qui a commencé la semaine dernière, **avec** les gens me pestering tout le temps.
✗	CNN	Je n' arrive pas à terminer l' un ou l' autre des travaux qui m' ont été confiés par mon nouveau gestionnaire qui a commencé la semaine dernière, **avec** des gens qui m' nuisent tout le temps.
✗	RNN	Je ne peux pas terminer le travail qui m' a été confié par mon nouveau gestionnaire qui a commencé la semaine dernière, **avec** des gens qui m' ont constamment blessé.
✗	Attention	Je ne peux terminer aucun des travaux que m' a confiés mon nouveau gestionnaire qui a commencé la semaine dernière, **avec** des gens qui m' ont ravageur tout le temps.
✓	Google	Je ne peux pas terminer le travail qui m'a été confié par mon nouveau responsable qui a commencé la semaine dernière, les gens me **harcelant** tout le temps.
✗	DeepL	Je n'arrive pas à finir les tâches qui m'ont été confiées par mon nouveau directeur, qui a commencé la semaine dernière, **avec** des gens qui me harcèlent tout le temps.
S10h	Source	I'm thinking of riding my brother's old bicycle to work and to the gym starting next week, **with** the price of gas on the rise.
	Ref	Je pense prendre l'ancien vélo de mon frère pour aller au travail et à la salle de gym à partir de la semaine prochaine, le prix de l'essence **étant** à la hausse.
✗	Hybrid SMT	La pensée I'm d' équitation mon frère 's vieux vélo au travail et au gymnase dès la semaine prochaine, **avec** le prix du gaz à la hausse.

✗	CNN	J' ai l' impression d' amener le vieux vélo de mon frère à travailler et à gymnase à partir de la semaine prochaine, **avec** le prix du gaz à la hausse.
✗	RNN	Je songe à monter la vieille bicyclette de mon frère pour travailler et au gymnase à partir de la semaine prochaine, **avec** le prix du gaz à la hausse.
✗	Attention	Je pense à faire monter la vieille bicyclette de mon frère au travail et au gymnase à partir de la semaine prochaine, **avec** la hausse du prix du gaz.
✗	Google	Je pense faire du vieux vélo de mon frère au travail et à la salle de sport à partir de la semaine prochaine, **avec** le prix de l'essence à la hausse.
✗	DeepL	Je pense à utiliser le vieux vélo de mon frère pour aller au travail et au gymnase à partir de la semaine prochaine, **avec** le prix de l'essence qui augmente.

"With" expressing a particular feeling or physical state

Possible solutions: Generally, *de* or *par* work in many cases. Other alternatives that sound more idiomatic (phrases like *atteint par* [*une maladie*]) are also accepted.

Short sentences		
S11a	Source	He was consumed **with** guilt after he broke his mother's favourite vase.
	Ref	Il était rongé **par** la culpabilité après avoir brisé le vase préféré de sa mère.
✗	Hybrid SMT	Il a été consommé **avec** la culpabilité après qu' il a battu sa mère 's vase favori.
✗	CNN	Il a été consommé **avec** culpabilité après avoir rompu son vase préféré de sa mère.
✗	RNN	Il a été consommé **avec** culpabilité après avoir brisé la vase préférée de sa mère.
✗	Attention	Il a été consommé **avec** culpabilité après avoir brisé la vase préférée de sa mère.
✓	Google	Il a été rongé **par** la culpabilité après avoir brisé le vase préféré de sa mère.
✓	DeepL	Il était rongé **par** la culpabilité après avoir cassé le vase préféré de sa mère.

S11b	Source	Everyone could see that her eyes were still burning **with** hatred.
	Ref	Tout le monde pouvait voir que ses yeux étaient toujours remplis **de** haine.
✗	Hybrid SMT	Tout le monde peut voir que ses yeux étaient encore brûler **avec** la haine.
✗	CNN	Tout le monde pouvait voir que ses yeux allaient encore brûler **avec** la haine.
✗	RNN	Tout le monde pouvait voir que ses yeux brûlaient encore **avec** la haine.
✗	Attention	Tout le monde pouvait voir que ses yeux brûlaient encore **avec** la haine.
✓	Google	Tout le monde pouvait voir que ses yeux brûlaient encore **de** haine.
✓	DeepL	Tout le monde pouvait voir que ses yeux brûlaient encore **de** haine.
S11c	Source	Still beaming **with** joy, she ran home to tell her parents the good news.
	Ref	Toujours rayonnante **de** joie, elle a couru jusqu'à la maison pour annoncer la bonne nouvelle à ses parents.
✓	Hybrid SMT	Toujours radieux **de** joie, elle dirigeait la maison de dire à ses parents la bonne nouvelle.
✓*	CNN	Toujours **avec** joie, elle a couru chez elle pour dire à ses parents la bonne nouvelle.
✗	RNN	Toujours **à** la joie, elle s' est rendue chez elle pour dire à ses parents les bonnes nouvelles.
✓*	Attention	Toujours **avec** joie, elle s' est rendue chez elle pour dire à ses parents les bonnes nouvelles.
✓	Google	Toujours radieuse **de** joie, elle a couru à la maison pour annoncer la bonne nouvelle à ses parents.
✓	DeepL	Toujours rayonnante **de** joie, elle a couru à la maison pour annoncer la bonne nouvelle à ses parents.
S11d	Source	That medication is not recommended for use in patients **with** hypothermia.
	Ref	Ce médicament n'est pas recommandé chez les patients **souffrant d'**hypothermie.
✓	Hybrid SMT	Ce médicament n' est pas recommandé chez les patients **atteints d'** hypothermie.
✓	CNN	Ce médicament n' est pas recommandé pour les patients **souffrant d'** hypothermie.

✓	RNN	Ce médicament n' est pas recommandé chez les patients **souffrant d'** hypothermie.
✓	Attention	Ce médicament n' est pas recommandé pour les patients **souffrant d'** hypothermie.
✓	Google	Ce médicament n'est pas recommandé chez les patients **souffrant d'**hypothermie.
✓	DeepL	Ce médicament n'est pas recommandé pour les patients **souffrant d'**hypothermie.
Long sentences		
S11e	Source	He was consumed **with** guilt, a feeling he had not felt since he accidentally ruined his sister's sweater in the dryer, after he broke his mother's favourite vase.
	Ref	Il était rongé **par** la culpabilité, un sentiment qu'il n'avait pas senti depuis qu'il avait accidentellement ruiné le pull de sa sœur dans la sécheuse, après avoir brisé le vase préféré de sa mère.
✗	Hybrid SMT	Il a été consommé **avec** la culpabilité, un sentiment qu' il n' avait pas ressenti depuis qu' il a ruiné sa sœur chandail 's accidentellement dans la sécheuse, après qu' il a battu sa mère 's vase favori.
✗*	CNN	Il a été consommé **de** culpabilité, un sentiment qu' il n' avait pas ressenti depuis qu' il avait accidentellement ruiné son chandail de la sœur dans la sécheuse, après avoir rompu sa mère "la vase préférée.
✗	RNN	Il a été consommé **avec** culpabilité, un sentiment qu' il n' avait pas ressenti depuis qu' il a accidentellement ruiné le chandail de sa soeur dans la sécheuse.
✗	Attention	Il a été consommé **avec** culpabilité, un sentiment qu' il n' avait pas ressenti puisqu' il a accidentellement ruiné le chandail de sa sœur dans le séchoir.
✓	Google	Il était rongé **par** la culpabilité, un sentiment qu'il n'avait pas ressenti depuis qu'il avait accidentellement ruiné le pull de sa sœur dans la sécheuse, après avoir cassé le vase préféré de sa mère.
✓	DeepL	Il était rongé **par** la culpabilité, un sentiment qu'il n'avait pas ressenti depuis qu'il avait accidentellement abîmé le pull de sa sœur dans le sèche-linge, après avoir cassé le vase préféré de sa mère.

S11f	Source	Although she claimed she had moved on and forgiven her family, everyone could see that her eyes still burned **with** a hatred so deep that no one dared to approach her.
	Ref	Bien qu'elle dise être passée à autre chose et avoir pardonné sa famille, tout le monde pouvait voir que ses yeux étaient toujours remplis **d'**une haine si profonde qu'on n'osait l'approcher.
✗	Hybrid SMT	Bien qu' elle prétend qu' elle avait évolué et pardonné sa famille, tout le monde peut voir que ses yeux brûlent encore **avec** une haine si profond que personne n' a osé son approche.
✗	CNN	Bien qu' elle prétend qu' elle avait déménagé et pardonné à sa famille, tout le monde pouvait voir que ses yeux étaient toujours brûlée **avec** une haine si profonde que personne n' a osé l' approcher.
✗	RNN	Bien qu' elle prétendait qu' elle avait déménagé et pardonné sa famille, tout le monde pouvait voir que ses yeux brûlaient encore **avec** une haine si profonde que personne n' osait l' approcher.
✗	Attention	Bien qu' elle ait affirmé qu' elle avait déménagé et pardonné sa famille, tout le monde pouvait voir que ses yeux étaient encore brûlés **avec** une haine tellement profonde que personne n' a osé.
✓	Google	Même si elle affirmait qu'elle avait évolué et pardonné à sa famille, tout le monde pouvait voir que ses yeux brûlaient encore **d'**une haine si profonde que personne n'osait l'approcher.
✓	DeepL	Bien qu'elle ait affirmé avoir tourné la page et pardonné à sa famille, tout le monde pouvait voir que ses yeux brûlaient encore **d'**une haine si profonde que personne n'osait l'approcher.
S11g	Source	Still beaming **with** joy, she ran home to tell her parents the good news, unaware that they had left home earlier to go visit her sick aunt.
	Ref	Toujours rayonnante **de** joie, elle a couru jusqu'à la maison pour annoncer la bonne nouvelle à ses parents, sans savoir qu'ils étaient partis un peu plus tôt pour aller rendre visite à sa tante malade.

✓	Hybrid SMT	Toujours radieux **de** joie, elle dirigeait la maison de dire à ses parents la bonne nouvelle, ignorant qu' ils avaient quitté la Maison plus tôt pour aller rendre visite à sa tante malade.
✗	CNN	Toujours à faisceau **avec** joie, elle a couru chez elle pour dire à ses parents les bonnes nouvelles, ignorées qu' ils avaient quitté la maison plus tôt pour aller visiter sa tante malade.
✗	RNN	Toujours **à** la joie, elle s' est rendue à la maison pour dire à ses parents les bonnes nouvelles, ignorant qu' ils avaient quitté la maison plus tôt pour aller visiter sa tante malade.
✓*	Attention	Toujours **avec** joie, elle s' est rendue chez elle pour dire à ses parents les bonnes nouvelles, sans savoir qu' ils avaient quitté la maison plus tôt pour aller visiter sa tante malade.
✓	Google	Toujours radieuse **de** joie, elle a couru à la maison pour annoncer la bonne nouvelle à ses parents, ignorant qu'ils avaient quitté la maison plus tôt pour rendre visite à sa tante malade.
✓	DeepL	Toujours rayonnante **de** joie, elle a couru à la maison pour annoncer la bonne nouvelle à ses parents, ignorant qu'ils avaient quitté la maison plus tôt pour aller rendre visite à sa tante malade.
S11h	Source	That medication has been observed to cause severe side effects including liver damage and is not recommended for use in patients **with** thyroid problems or severe hypothermia.
	Ref	Ce médicament cause de graves effets secondaires, y compris des dommages au foie, et n'est pas recommandé chez les patients **souffrant de** problèmes thyroïdiens ou d'hypothermie grave.
✓	Hybrid SMT	Ce médicament a été observé pour causer des effets secondaires graves, y compris des dommages au foie et n' est pas recommandé chez les patients **présentant des** problèmes thyroïdiens ou d' hypothermie grave.
✗	CNN	On a observé que ce médicament provoque des effets secondaires graves, y compris les dommages au foie et n' est pas recommandé pour les patients **avec** des problèmes thyroïdiens ou d' hypothermie grave.
✓	RNN	Ce médicament a des effets secondaires graves, y compris des lésions hépatiques et n' est pas recommandé chez les patients **souffrant de** problèmes thyroïdiens ou d' hypothermie grave.

✓	Attention	On a observé que ce médicament cause de graves effets secondaires, y compris des lésions hépatiques, et il n' est pas recommandé pour les patients **souffrant de** troubles thyroïdiens ou d' hypothermie.
✓	Google	Il a été observé que ce médicament provoque des effets indésirables graves, notamment des lésions hépatiques, et son utilisation n'est pas recommandée chez les patients **souffrant de** problèmes thyroïdiens ou d'hypothermie sévère.
✓	DeepL	Il a été observé que ce médicament provoque de graves effets secondaires, notamment des lésions hépatiques, et qu'il n'est pas recommandé aux patients **souffrant de** problèmes de thyroïde ou d'hypothermie grave.

"With" meaning "in spite of"

Possible solutions: *malgré*, *en dépit de*, etc.

Short sentences		
"with" at the start of the sentence		
S12a	Source	**With** that bad weather, they still went on their daily walk.
	Ref	Malgré ce mauvais temps, ils sont allés faire leur promenade quotidienne.
✗	Hybrid SMT	**Avec** ce mauvais temps, ils sont allés encore sur leur marche quotidienne.
✗	CNN	**Avec** ce mauvais temps, ils ont toujours fait leurs promenades quotidiennes.
✗	RNN	**Avec** ce mauvais temps, ils ont continué à marcher quotidiennement.
✗	Attention	**Avec** ce mauvais temps, ils ont continué à marcher tous les jours.
✗	Google	**Avec** ce mauvais temps, ils continuaient leur promenade quotidienne.
✗	DeepL	**Avec** ce mauvais temps, ils continuaient à faire leur promenade quotidienne.
S12b	Source	**With** all his debts, he not only bought a car, but also got leather seats
	Ref	**Malgré** toutes ses dettes, il a non seulement acheté une voiture, mais a aussi choisi des sièges en cuir.

✗	Hybrid SMT	**Avec** toutes ses dettes, il a non seulement acheté une voiture, mais a également obtenu des sièges en cuir.
✗	CNN	**Avec** toutes ses dettes, il a non seulement acheté une voiture, mais aussi des sièges en cuir.
✗	RNN	**Avec** toutes ses dettes, il n ’ a pas seulement acheté une voiture, mais a également obtenu des sièges en cuir.
✗	Attention	**Avec** toutes ses dettes, il a non seulement acheté une voiture, mais aussi obtenu des sièges en cuir.
✗	Google	**Avec** toutes ses dettes, il a non seulement acheté une voiture, mais aussi des sièges en cuir.
✗	DeepL	**Avec** toutes ses dettes, il a non seulement acheté une voiture, mais il a aussi obtenu des sièges en cuir.
“with” mid-sentence		
S12c	Source	I still think we should launch the project, **with** all the risks it entails.
	Ref	Je pense tout de même qu’on devrait lancer le projet **en dépit des** risques qu’il comporte.
✗	Hybrid SMT	Je pense que nous devrions lancer le projet, **avec** tous les risques que cela implique.
✗	CNN	Je pense toujours que nous devrions lancer le projet, **avec** tous les risques qu' il comporte.
✗	RNN	Je pense toujours que nous devrions lancer le projet, **avec** tous les risques qu' il comporte.
✗	Attention	Je pense toujours que nous devrions lancer le projet, **avec** tous les risques qu' il comporte.
✗	Google	Je pense toujours qu'il faut lancer le projet, **avec** tous les risques qu'il comporte.
✗	DeepL	Je pense toujours que nous devrions lancer le projet, **avec** tous les risques qu'il comporte.
S12d	Source	The professor let those two students pass, **with** their numerous absences.
	Ref	Le professeur a accordé la note de passage à ces deux étudiants, **en dépit de** leurs nombreuses absences.
✗	Hybrid SMT	Le professeur laissons ces deux étudiants passent, **avec** leurs nombreuses absences.
✗	CNN	Le professeur a permis à ces deux élèves de réussir, **avec** leurs nombreuses absences.
✗	RNN	Le professeur laisse passer ces deux étudiants, **avec** leurs nombreuses absences.

✗	Attention	Le professeur laisse passer ces deux étudiants, **avec** leurs nombreuses absences.
✗	Google	Le professeur a laissé passer ces deux étudiants, **avec** leurs nombreuses absences.
✗	DeepL	Le professeur a laissé passer ces deux étudiants, **avec** leurs nombreuses absences.
Long sentences		
"with" at the start of the sentence		
S12e	Source	**With** the unexpectedly cold and cloudy weather that everyone else was complaining about, they still went on their daily walk, only wearing a light jacket.
	Ref	**Malgré** ce temps étonnament froid, nuageux et dont tout le monde se plaignait, ils sont allés faire leur promenade quotidienne en ne portant qu'une veste légère.
✗	Hybrid SMT	**Avec** le temps froid et nuageux de façon inattendue que tout le monde se plaignait, ils sont allés encore sur leur marche quotidienne, ne portait une veste légère.
✗	CNN	**Avec** le temps froid et nuageux inattendus que tout le monde s' est plaint, ils sont toujours allés sur leur promenade quotidienne, portant seulement une veste légère.
✗	RNN	**Avec** le temps inattendu et nuageux que tous les autres se plaignent, ils ont continué à marcher quotidiennement, portant uniquement une veste légère.
✗	Attention	**Avec** le temps inattendument froid et nuageux dont tout le monde se plaint, ils passèrent toujours leur marche quotidienne, ne portant qu' un gilet léger.
✗	Google	**Avec** le temps froid et nuageux inattendu dont tout le monde se plaignait, ils continuaient leur promenade quotidienne, ne portant qu'une veste légère.
✗	DeepL	**Avec** le temps froid et nuageux inattendu dont tout le monde se plaignait, ils ont continué leur promenade quotidienne, ne portant qu'une veste légère.
S12f	Source	**With** all his debts, he not only bought a new car from the dealership with his credit card, he also got leather seats and a panoramic sunroof.
	Ref	**Malgré** toutes ses dettes, il a non seulement acheté une nouvelle voiture du concessionnaire avec sa carte de crédit, mais a aussi choisi des sièges en cuir et un toit ouvrant panoramique.

✗	Hybrid SMT	**Avec** toutes ses dettes, non seulement il a acheté une nouvelle voiture du concessionnaire avec sa carte de crédit, il a également obtenu des sièges en cuir et un toit panoramique.
✗	CNN	**Avec** toutes ses dettes, il a non seulement acheté une nouvelle voiture du concessionnaire avec son crédit, il a également obtenu des sièges en cuir et un toit de soleil panoramique.
✗	RNN	**Avec** toutes ses dettes, il n' a pas seulement acheté une nouvelle voiture du concessionnaire avec sa carte de crédit, il a également obtenu des sièges en cuir et un toit panoramique.
✗	Attention	**Avec** toutes ses dettes, il n' a pas seulement acheté une nouvelle voiture du concessionnaire avec sa carte de crédit, il a également obtenu des sièges en cuir et un toit de soleil panoramique.
✗	Google	**Avec** toutes ses dettes, il a non seulement acheté une nouvelle voiture au concessionnaire avec sa carte de crédit, mais il a également obtenu des sièges en cuir et un toit ouvrant panoramique.
✗	DeepL	**Avec** toutes ses dettes, il a non seulement acheté une nouvelle voiture chez le concessionnaire avec sa carte de crédit, mais il a également obtenu des sièges en cuir et un toit ouvrant panoramique.
"with" mid-sentence		
S12g	Source	I still think we should go forward with that project, **with** all the risks it entails and considering all the efforts that would be required to execute it.
	Ref	Je pense tout de même qu'on devrait lancer le projet **en dépit des** risques qu'il comporte et compte tenu de tous les efforts qui seront nécessaire pour son exécution.
✗	Hybrid SMT	Je pense que nous devrions aller de l' avant avec ce projet, **avec** tous les risques que cela implique et en tenant compte de tous les efforts qui seraient nécessaires pour l' exécuter.
✗	CNN	Je pense toujours que nous devrions aller de l' avant avec ce projet, **avec** tous les risques qu' il comporte et prend en considération tous les efforts qui seraient nécessaires pour l' exécuter.

✗	RNN	Je pense toujours que nous devrions aller de l' avant avec ce projet, **avec** tous les risques qu' il comporte et considérer tous les efforts qui seraient nécessaires pour l' exécuter.
✗	Attention	Je pense encore que nous devrions aller de l' avant avec ce projet, **avec** tous les risques qu' il comporte et en tenant compte de tous les efforts qui seraient nécessaires pour l' exécuter.
✗	Google	Je pense toujours que nous devrions aller de l'avant avec ce projet, **avec** tous les risques qu'il comporte et en tenant compte de tous les efforts qui seraient nécessaires pour l'exécuter.
✗	DeepL	Je pense toujours que nous devrions aller de l'avant avec ce projet, **avec** tous les risques qu'il comporte et en considérant tous les efforts qui seraient nécessaires pour le réaliser.
S12h	Source	Since the school did not have any specific policy or rules regarding attendance, the professor let those two students pass the course, **with** their numerous absences.
	Ref	Comme l'école n'avait pas de politique ou de règlements spécifiques liés à la présence, le professeur a laissé ces deux étudiants réussir le cours, **en dépit de** leurs nombreuses absences.
✗	Hybrid SMT	Étant donné que l' école n' avait aucune politique spécifique ou les règles concernant les présences, le professeur laissons ces deux étudiants passent le cap, **avec** leurs nombreuses absences.
✗	CNN	Étant donné que l ' école n ' avait pas de politique ou de règles particulières en ce qui concerne la présence, le professeur laissait ces deux Les étudiants passent le cours, **avec** leurs nombreuses absences.
✗	RNN	Étant donné que l' école n' avait pas de politique ou de règles spécifiques concernant la fréquentation, le professeur laissait ces deux étudiants passer le cours, **avec** leurs nombreuses absences.
✗	Attention	Comme l' école n' avait pas de politique ou de règles spécifiques concernant la fréquentation, le professeur a permis à ces deux étudiants de réussir le cours, **avec** leurs nombreuses absences.

✗	Google	Comme l'école n'avait pas de politique ou de règles spécifiques concernant l'assiduité, le professeur a laissé ces deux étudiants réussir le cours, **avec** leurs nombreuses absences.
✗	DeepL	Comme l'école n'avait pas de politique ou de règles spécifiques concernant l'assiduité, le professeur a laissé ces deux étudiants suivre le cours, **avec** leurs nombreuses absences.

Homographs

Multiple words in English translated as one word in French
Possible solution: find a turn of phrase or a synonym to avoid a repetition

Short sentences		
S13a	Source	The company bought new **desks** for all of their **offices**.
	Ref	La compagnie a acheté de nouveaux **bureaux** pour tous ses **édifices**.
✗	Hybrid SMT	L' entreprise a acheté de nouveaux **bureaux** pour l' ensemble de leurs **bureaux**.
✗	CNN	L' entreprise a acheté de nouveaux **bureaux** pour l' ensemble de leurs **bureaux**.
✗	RNN	L' entreprise a acheté de nouveaux **bureaux** pour tous leurs **bureaux**.
✗	Attention	L' entreprise a acheté de nouveaux **bureaux** pour tous ses **bureaux**.
✗	Google	L'entreprise a acheté de nouveaux **bureaux** pour tous ses **bureaux**.
✗	DeepL	L'entreprise a acheté de nouveaux **bureaux** pour tous ses **bureaux**.
S13b	Source	While I **appreciate** the comments people leave on my videos, I don't particularly **enjoy** them.
	Ref	Bien que **j'apprécie** les commentaires que les gens me laissent sur mes vidéos, je ne les **aime** pas particulièrement.
✓	Hybrid SMT	Bien que j' **apprécie** les commentaires des gens laissent sur mes vidéos, je don't particulièrement en **profiter**.
✗	CNN	Bien que j' **apprécie** les commentaires que les gens partent sur mes vidéos, je n' **apprécie** pas particulièrement C ' est ce qu ' ils ont fait.

✓	RNN	Bien que j' **apprécie** les commentaires que les gens partent sur mes vidéos, je ne les **aime** pas particulièrement.
✗	Attention	Bien que j' **apprécie** les commentaires que les gens partent sur mes vidéos, je ne les **apprécie** pas particulièrement.
✗	Google	Bien que j'**apprécie** les commentaires que les gens laissent sur mes vidéos, je ne les **apprécie** pas particulièrement.
✗	DeepL	Bien que j'**apprécie** les commentaires que les gens laissent sur mes vidéos, je ne les **apprécie** pas particulièrement.
S13c	Source	He didn't exactly know whether what he felt was **like** or **love**.
	Ref	Il ne savait pas exactement s'il ressentait de **l'appréciation** ou de **l'amour**.
✗	Hybrid SMT	Il a didn't exactement de savoir si ce qu' il croyait était **comme** ou l' **amour**.
✗	CNN	Il ne savait pas exactement si ce qu' il s' était senti **comme** ou l' **amour**.
✗	RNN	Il ne savait pas exactement si ce qu' il sentait était **aimé** ou **aimé**.
✗	Attention	Il ne savait pas exactement si ce qu' il pensait était **comme** ou si il **aimait**.
✗	Google	Il ne savait pas exactement si ce qu'il ressentait était **comme** ou si c'était de l'**amour**.
✓	DeepL	Il ne savait pas exactement si ce qu'il ressentait était de l'**amour** ou **non**.
S13d	Source	The wine was delicious: **cool** and **fresh** and fruity.
	Ref	Le vin était délicieux : **frais**, d'un goût **rafraîchissant** et fruité.
✗*	Hybrid SMT	Le vin était délicieux : **cool** et **frais** et fruité.
✗	CNN	Le vin était délicieux: **frais** et **frais**.
✗	RNN	Le vin était délicieux: **frais** et **frais** et fruité.
✓*	Attention	Le vin était délicieux: **frais** et fruité.
✗	Google	Le vin était délicieux: **frais** et **frais** et fruité.
✓*	DeepL	Le vin était délicieux : **frais** et fruité.

Long sentences		
S13e	Source	With this year’s new budget and with new employees joining in the coming weeks, the company bought new chairs and new **desks** for their **offices**.
	Ref	Avec le nouveau budget de cette année et l’arrivée de nouveaux employés dans les semaines à venir, la compagnie a acheté de nouvelles chaises et de nouveaux **bureaux** pour tous ses **immeubles**.
✗	Hybrid SMT	Avec cette année 's nouveau budget et avec les nouveaux employés dans les semaines à venir, la compagnie a acheté de nouveaux présidents et de nouveaux **bureaux** pour leurs **bureaux**.
✗	CNN	Avec le nouveau budget de cette année et les nouveaux employés se joignent dans les prochaines semaines, L' entreprise a acheté de nouvelles chaises et de nouveaux **bureaux** pour leurs **bureaux**.
✗	RNN	Avec le nouveau budget de cette année et avec l' arrivée des nouveaux employés dans les semaines à venir, l' entreprise a acheté de nouveaux chaises et de nouveaux **bureaux** pour leurs **bureaux**.
✗	Attention	Grâce au nouveau budget de cette année et à l' arrivée de nouveaux employés dans les semaines à venir, l' entreprise a acheté de nouveaux chaises et de nouveaux **bureaux** pour ses **bureaux**.
✗	Google	Avec le nouveau budget de cette année et l'arrivée de nouveaux employés dans les semaines à venir, la société a acheté de nouvelles chaises et de nouveaux **bureaux** pour ses **bureaux**.
✗	DeepL	Avec le nouveau budget de cette année et l'arrivée de nouveaux employés dans les semaines à venir, la société a acheté de nouvelles chaises et de nouveaux **bureaux** pour ses **bureaux**.
S13f	Source	Because they are thoughtful and I understand where they are coming from, I **appreciate** but don’t particularly **enjoy** the comments you leave on my videos about weight loss.
	Ref	Puisqu’ils sont attentionnés et que je comprends d’où ils viennent, **j’apprécie** mais je **n’aime** pas particulièrement les commentaires que vous me laissez sur mes vidéos portant sur la perte de poids.

✓	Hybrid SMT	Parce qu' ils sont réfléchis et je comprends où ils arrivent, j' **apprécie** mais don't **aiment** particulièrement les commentaires que vous quittez sur mes vidéos sur la perte de poids.
✗	CNN	Parce qu' ils sont réfléchis et je comprends où ils viennent, j' **apprécie**, mais pas ». Vous **apprécierez** particulièrement les commentaires que vous partez sur mes vidéos sur la perte de poids.
✗	RNN	Parce qu' ils sont réfléchis et je comprends d' où ils viennent, j' **apprécie** mais je n' **apprécie** pas particulièrement les commentaires que vous laisse sur mes vidéos sur la perte de poids.
✗	Attention	Parce qu' ils sont réfléchis et que je comprends d' où ils viennent, j' **apprécie** mais n' **apprécie** pas particulièrement les commentaires que vous partez sur mes vidéos sur la perte de poids.
✗	Google	Parce qu'ils sont réfléchis et que je comprends d'où ils viennent, j'**apprécie** mais j'**apprécie** particulièrement les commentaires que vous laissez sur mes vidéos sur la perte de poids.
✓	DeepL	Parce qu'ils sont bien pensés et que je comprends leurs origines, j'**apprécie**, mais pas particulièrement, les commentaires que vous laissez sur mes vidéos sur la perte de poids.
S13g	Source	He had just gotten out of a serious relationship that had lasted four years and he didn't exactly know whether what he felt now was **like**— what he would feel for any good friend he had grown to rely on—or **love**.
	Ref	Il venait de sortir d'une relation sérieuse qui avait duré quatre ans et ne savait pas exactement si ce qu'il ressentait en ce moment était de **l'appréciation**, un sentiment qu'il ressentirait pour tout ami sur qui il aurait appris à se fier, ou de **l'amour**.
✗	Hybrid SMT	Il avait obtenu une relation sérieuse qui a duré quatre ans et il didn't exactement de savoir si ce qu' il croyait maintenant était **comme** - ce qu' il se sentirait pour un bon ami, il avait grimpé à compter sur - ou l' **amour**.

✗	CNN	Il venait tout juste de s' éloigner d' une relation sérieuse qui avait duré quatre ans. sait exactement si ce qu' il **semblait** maintenant, ce qu' il se sentirait pour un bon ami s' étaient développés pour s' en remettre - ou à l' **amour**.
✗*	RNN	Il venait tout juste de sortir d' une relation sérieuse qui avait duré quatre ans et il ne savait pas exactement si ce qu' il ressentait maintenant était **comme** - ce qu' il ressentait pour tout bon ami.
✗*	Attention	Il venait de sortir d' une relation sérieuse qui avait duré quatre ans et il ne savait pas exactement si ce qu' il pensait aujourd' hui était - ce qu' il ressentait pour un bon ami.
✗*	Google	Il venait juste de sortir d'une relation sérieuse qui avait duré quatre ans et il ne savait pas exactement si ce qu'il ressentait maintenant était - ce qu'il ressentirait pour tout bon ami sur lequel il avait appris à compter - ou s'il **aimait**.
✗*	DeepL	Il venait de sortir d'une relation sérieuse qui avait duré quatre ans et il ne savait pas exactement si ce qu'il ressentait maintenant était comme - ce qu'il ressentirait pour tout bon ami sur lequel il avait appris à compter ou qu'il **aimait**.
S13h	Source	The wine we tasted at the first stop of our wine tour was delicious: **cool**, **fresh**, and fruity, with notes of vanilla and sage, aged in old oak barrels.
	Ref	Le vin auquel nous avons goûté au premier arrêt de notre visite de vignobles était délicieux : **frais,** d'un goût **rafraichissant**, fruité, avec des notes de vanille et de sauge et vieilli en fûts de chêne.
✗*	Hybrid SMT	Le vin que nous avons goûté au premier arrêt de notre tournée de vin était délicieux : **cool**, **frais** et fruité, avec des notes de vanille et de sauge, âgés de vieux barils de chêne.
✗	CNN	Le vin que nous avons goûté au premier arrêt de notre tour de vin était délicieux: **frais**, **frais**, **frais**, et fruité, avec des notes de vanille et de sage, âgés fûts de chêne.
✗	RNN	Le vin que nous avons dégusté au premier arrêt de notre tour de vin était délicieux: **frais**, **frais**, fruité, avec des notes de vanille et de sauge, vieilli dans des fûts de chêne.
✗	Attention	Le vin que nous avons dégusté à la première étape de notre tournée de vin était délicieux: **frais**, **frais** et fruité, avec notes de vanille et de sage, vieux barils de chêne.

✗	Google	Le vin que nous avons goûté au premier arrêt de notre visite des vins était délicieux: **frais**, **frais** et fruité, avec des notes de vanille et de sauge, vieilli dans de vieux fûts de chêne.
✓	DeepL	Le vin que nous avons dégusté lors de la première étape de notre circuit était délicieux : **frais**, **puissant** et fruité, avec des notes de vanille et de sauge, vieilli dans de vieux fûts de chêne.

Appendix B: Challenge set – Syntactic difficulties

Scope

Scope of modifiers

Short sentences		
S14a	Source	The only **Australian wine lovers** I know are coincidentally also Australian.
	Ref	Les seuls **amateurs de vin australien** que je connaisse sont, par coïncidence, eux aussi australiens.
✓	Hybrid SMT	Seuls les **amateurs de vin australien** je sais sont aussi incidemment australien.
✗	CNN	Les seuls **amateurs de vin australiens** que je connais est aussi l' Australien.
✗	RNN	Les seuls **amateurs de vin australiens** que je connais sont également australiens.
✗	Attention	Les seuls **amateurs de vin australiens** que je connais sont également australiens.
✓*	Google	Les seuls **amateurs de vins australiens** que je connaisse sont également australiens.
✗	DeepL	Les seuls **amateurs de vin australiens** que je connaisse sont, par coïncidence, également australiens.
S14b	Source	Don't tell me you lost the **dehydrated dog treat** I just got from the store.
	Ref	Ne me dites pas que vous avez perdu la **gâterie déshydratée pour chien** que je viens d'acheter au magasin.
✗	Hybrid SMT	Don't me dire que vous avez perdu le **chien déshydraté traiter** je viens de recevoir du magasin.
✗	CNN	Ne me dites pas que vous avez perdu le **chien déshydraté** que j' ai reçu du magasin..
✗	RNN	Ne me dites pas que vous avez perdu le **traitement des chiens déshydratés** que je viens juste d' obtenir du magasin.
✓	Attention	Ne me dites pas que vous avez perdu le **traitement déshydraté des chiens** que je viens de recevoir du magasin.
✓	Google	Ne me dites pas que vous avez perdu la **friandise déshydratée pour chien** que je viens de recevoir du magasin.
✗	DeepL	Ne me dites pas que vous avez perdu la **friandise pour chien déshydraté** que je viens de recevoir du magasin.

S14c	Source	Our store has a **new arrivals window display** and one showcasing our most popular items.
	Ref	Notre magasin a une **vitrine pour ses nouveautés** et une mettant en vedette nos articles les plus populaires.
✗	Hybrid SMT	Notre magasin a une **nouvelle fenêtre d' affichage des arrivées** et une vitrine de nos articles les plus populaires.
✗	CNN	Notre magasin dispose **d' une nouvelle fenêtre d' arrivées** et une vitrine de nos articles les plus populaires.
✗	RNN	Notre magasin dispose d' une **nouvelle fenêtre d' arrivée** et l' un de nos articles les plus populaires.
✗	Attention	Notre magasin dispose d' une **nouvelle fenêtre d' arrivées** et d' une fenêtre présentant nos articles les plus populaires.
✓	Google	Notre magasin a une **vitrine pour les nouveaux arrivants** et une présentant nos articles les plus populaires.
✓	DeepL	Notre magasin a une **vitrine pour les nouveaux arrivants** et une autre pour nos articles les plus populaires.
S14d	Source	If you think you have trouble cooking asparagus, listen to my **abnormally crisp asparagus story**.
	Ref	Si vous pensez avoir du mal à faire cuire des asperges, écoutez mon **histoire d'asperges anormalement croustillantes**.
✗	Hybrid SMT	Si vous pensez que vous avez de la difficulté à la cuisson des asperges, écoutez **mon histoire anormalement croquante des asperges**.
✗*	CNN	Si vous croyez avoir de la difficulté à faire cuire l' asperge, écoutez mes **anormalement criants**: L' histoire de l' épereur.
✓	RNN	Si vous pensez avoir de la difficulté à cuisiner des asperges, écoutez mon **histoire d' asperges anormalement criantes**.
✗	Attention	Si vous pensez avoir de la difficulté à cuire des asperges, écoutez mon **histoire anormalement cristalline d' asperges**.
✓	Google	Si vous pensez avoir du mal à cuire des asperges, écoutez mon **histoire d'asperges anormalement croquantes**.
✓	DeepL	Si vous pensez que vous avez du mal à cuire les asperges, écoutez mon **histoire d'asperges anormalement croquantes**.

Long sentences		
S14e	Source	I don’t know many people who would choose Australian wine over French wine; in fact, the only **Australian wine lovers** I know are coincidentally also Australian.
	Ref	Je connais peu de gens qui choisiraient du vin australien plutôt que du vin français, d’ailleurs, les seuls **amateurs de vin australien** que je connaisse sont, par coïncidence, eux aussi australiens.
✓	Hybrid SMT	Je sais don't de nombreuses personnes qui choisiraient le vin australien sur le vin français; en fait, les seuls **amateurs de vin australien** je sais sont aussi incidemment australien.
✗	CNN	Je ne connais pas beaucoup de gens qui choisiraient le vin australien sur le vin français; en fait, les seuls **amateurs de vin australiens** que je connais sont aussi d' ailleurs aussi australiens.
✗	RNN	Je ne connais pas beaucoup de gens qui choisiraient le vin australien au-dessus du vin français; en fait, les seuls **amateurs de vin australiens** que je connais sont aussi australiens.
✗	Attention	Je ne connais pas beaucoup de gens qui choisiraient le vin australien sur le vin français; en fait, les seuls **amateurs de vin australiens** que je connais sont également australiens.
✓*	Google	Je ne connais pas beaucoup de gens qui choisiraient le vin australien plutôt que le vin français; en fait, les seuls **amateurs de vins australiens** que je connaisse sont également australiens.
✓	DeepL	Je ne connais pas beaucoup de gens qui préféreraient le vin australien au vin français ; en fait, les seuls **amateurs de vin australien** que je connaisse sont, par coïncidence, également australiens.
S14f	Source	Don’t tell me you lost the very expensive, chicken-flavoured, certified organic, delectable **dehydrated dog treat** that I had finally just gotten from the store.
	Ref	Ne me dites pas que vous avez perdu la délectable **gâterie déshydratée pour chien**, très dispendieuse, au goût de poulet et certifiée biologique que je venais finalement d’acheter au magasin.

✗	Hybrid SMT	Don't me dire que vous avez perdu le très cher, chicken-flavoured, certifié biologique, délectable **chien déshydratés traiter** que j' avais finalement obtenu simplement du magasin.
✗	CNN	Ne me dites pas que vous avez perdu la saveur de poulet très chère, **un chien,** certifié biologique, **déshydraté**, que j' avais finalement obtenu de l' aide. Le magasin s' en sert.
✗*	RNN	Ne me dites pas que vous avez perdu le très chère, de poulet certifié organique, certifié, **déshydratable et déshydraté** que j' avais finalement venu du magasin.
✗*	Attention	Ne me dites pas que vous avez perdu le **traitement de chien déshydraté** très coûteux, à saveur de poulet, certifié biologique et délectable que j' avais enfin pris du magasin.
✓	Google	Ne me dites pas que vous avez perdu la très chère **gâterie déshydratée pour chiens** à saveur de poulet, certifiée biologique et délectable que je venais de recevoir du magasin.
✗	DeepL	Ne me dites pas que vous avez perdu la très chère **friandise pour chien déshydraté**, au goût de poulet, certifiée biologique et délectable, que je venais enfin d'acheter au magasin.
S14g	Source	The manager said our store should have a **new arrivals window display** to show the latest trends and another one showcasing our most popular items.
	Ref	Le gérant a dit que notre magasin devrait avoir **une vitrine avec ses nouveautés** pour montrer les dernières tendances et une autre mettant en vedette nos articles les plus populaires.
✓	Hybrid SMT	Le gestionnaire a déclaré notre magasin devrait avoir une **vitrine de nouveaux arrivants** pour montrer les dernières tendances et une autre présentant nos articles les plus populaires.
✗	CNN	Le gestionnaire a dit que notre magasin devrait afficher une **nouvelle fenêtre d' arrivée** pour montrer les dernières tendances et un autre Vous pouvez présenter nos articles les plus populaires.
✗	RNN	Le gestionnaire a indiqué que notre magasin devrait disposer d' une **nouvelle fenêtre d' arrivée** pour afficher les dernières tendances et une autre présentant nos articles les plus populaires.

✗	Attention	Le gestionnaire a dit que notre magasin devrait afficher une **nouvelle fenêtre d' arrivée** pour montrer les dernières tendances et un autre pour présenter nos articles les plus populaires.
✗	Google	Le gérant a déclaré que notre magasin devrait avoir une **nouvelle vitrine des arrivées** pour afficher les dernières tendances et une autre présentant nos articles les plus populaires.
✓	DeepL	Le gérant a déclaré que notre magasin devrait avoir une **vitrine des nouveaux arrivages** pour montrer les dernières tendances et une autre présentant nos articles les plus populaires.
S14h	Source	If you think you have trouble cooking asparagus in olive oil, with garlic, listen to my **abnormally crisp asparagus story** that I like to tell everyone.
	Ref	Si vous pensez avoir du mal à faire cuire des asperges dans de l'huile d'olive, avec de l'ail, écoutez mon **histoire d'asperges anormalement croustillantes** que j'aime raconter à tout le monde.
✗	Hybrid SMT	Si vous pensez que vous avez de la difficulté des asperges de cuisson dans l' huile d' olive, à l' ail, écoutez mon **histoire anormalement croquante asperge** que j' aime dire à tout le monde.
✗	CNN	Si vous pensez avoir de la difficulté à faire cuire l' asperge dans l' huile d' olive, avec l' ail, écoutez à mon que j' aimerais dire **histoire anormalement critiquée** à tout le monde.
✗*	RNN	Si vous pensez avoir de la difficulté à cuisiner des asperges dans l' huile d' olive, avec de l' ail, écoutez mon **histoire d' asperges anormalement criant** que j' aime dire à tout le monde.
✗*	Attention	Si vous pensez avoir de la difficulté à cuire des asperges dans l' huile d' olive, avec l' ail, écoutez mon **histoire d' asperges anormalement cristé** que j' aime.
✓	Google	Si vous pensez avoir du mal à cuire des asperges dans l'huile d'olive, avec de l'ail, écoutez mon **histoire d'asperges anormalement croquantes** que j'aime raconter à tout le monde.

✓	DeepL	Si vous pensez que vous avez du mal à cuire des asperges dans de l'huile d'olive, avec de l'ail, écoutez mon **histoire d'asperges anormalement croquantes** que j'aime raconter à tout le monde.

Scope of conjunction

Short sentences		
S15a	Source	I always bring two **solar-powered calculators and pencils** to my exams.
	Ref	J'apporte toujours **deux calculatrices solaires et des crayons** à mes examens.
✓	Hybrid SMT	J' ai l' habitude de faire **deux calculatrices solaires et crayons** à mes examens.
✗	CNN	J' apporte toujours **deux calculateurs et crayons à l' énergie solaire**.
✗	RNN	J' apporte toujours **deux calculatrices et crayons à énergie solaire** à mes examens.
✗	Attention	J' apporte toujours **deux calculateurs et crayons à énergie solaire** à mes examens.
✓	Google	J'apporte toujours **deux calculatrices solaires et des crayons** à mes examens.
✓	DeepL	J'apporte toujours **deux calculatrices à énergie solaire et des crayons** pour mes examens.
S15b	Source	That cream can be used to hydrate **itchy skin or wood**.
	Ref	Cette crème peut être utilisée pour hydrater **la peau qui démange ou le bois**.
✗*	Hybrid SMT	Que la crème peut être utilisée pour hydrater **des démangeaisons de la peau ou du bois**.
✓	CNN	Cette crème peut être utilisée pour hydrater **la peau de démangeaison ou le bois**.
✗*	RNN	Cette crème peut être utilisée pour hydrater **la peau ou le bois**.
✓	Attention	Cette crème peut être utilisée pour hydrater **la peau d' itchy ou le bois**.
✗	Google	Cette crème peut être utilisée pour hydrater **la peau ou le bois qui démange**.
✓	DeepL	Cette crème peut être utilisée pour hydrater **la peau qui démange ou le bois**.

S15c	Source	I have many pets at home, including **fluffy dogs and geckos**.
	Ref	J'ai beaucoup d'animaux de compagnie à la maison, y compris des **chiens poilus et des geckos**.
✗	Hybrid SMT	J' ai beaucoup d' animaux de compagnie à la maison, y compris les **chiens et les geckos duveteuses**.
✗*	CNN	J' ai beaucoup d' animaux de compagnie à la maison, y compris des **chiens et des geckos**.
✗*	RNN	J' ai beaucoup d' animaux de compagnie à la maison, y compris les **chiens et les geckos**.
✗*	Attention	J' ai beaucoup d' animaux de compagnie à la maison, y compris des **chiens et des geckos**.
✗	Google	J'ai beaucoup d'animaux à la maison, y compris des **chiens et des geckos moelleux**.
✓	DeepL	J'ai de nombreux animaux domestiques à la maison, notamment des **chiens en peluche et des geckos**.
S15d	Source	I ordered **wood blinds and curtains** online because they were cheaper.
	Ref	J'ai commandé des **stores en bois et des rideaux** en ligne parce qu'ils coûtaient moins cher.
✗*	Hybrid SMT	J' ai ordonné de **stores et de rideaux** en ligne parce qu' ils étaient moins chers.
✗*	CNN	J' ai commandé des **stores et des rideaux** en ligne parce qu' ils étaient moins chers.
✓	RNN	J' ai commandé des **stores en bois et des rideaux** en ligne parce qu' ils étaient moins chers.
✗*	Attention	J' ai commandé des **stores et des rideaux** en ligne parce qu' ils étaient moins chers.
✗	Google	J'ai commandé des **stores et des rideaux en bois** en ligne car ils étaient moins chers.
✗	DeepL	J'ai commandé des **stores et des rideaux en bois** en ligne parce qu'ils étaient moins chers.
Long sentences		
S15e	Source	I always bring two **solar-powered calculators and pencils** to my exams, in case one of my classmates forgot theirs and needs to borrow them from me.
	Ref	J'apporte toujours deux **calculatrices solaires et des crayons** à mes examens, au cas où un de mes camarades de classe aurait oublié les siennes et aurait besoin de me les emprunter.

✓	Hybrid SMT	J' ai l' habitude de faire deux **calculatrices solaires et crayons** à mes examens, au cas où l' un de mes collègues a oublié les leurs et doit emprunter de moi.
✗	CNN	J ’ apporte toujours deux **calculatrices et crayons à l ’ énergie solaire** à mes examens, au cas où l' un de mes camarades de classe m' a oublié et a besoin de les emprunter.
✗	RNN	J' apporte toujours deux **calculatrices et crayons à énergie solaire** à mes examens, au cas où l' un de mes camarades de classe a oublié le leur et a besoin de les emprunter de moi.
✗	Attention	J' apporte toujours deux **calculateurs et crayons à énergie solaire** à mes examens, au cas où l' un de mes camarades de classe aurait oublié de me les emprunter.
✓	Google	J'apporte toujours **deux calculatrices solaires et des crayons** à mes examens, au cas où l'un de mes camarades de classe oublierait les leurs et aurait besoin de me les emprunter.
✓	DeepL	J'apporte toujours **deux calculatrices solaires et des crayons** à mes examens, au cas où un de mes camarades de classe aurait oublié les leurs et aurait besoin de me les emprunter.
S15f	Source	That all-in-one cream can be used to hydrate **itchy skin or wood**, including wood that is too dry for woodworking or furniture that need to be restored.
	Ref	Cette crème tout-en-un peut être utilisée pour hydrater **la peau qui démange ou le bois**, y compris le bois qui est trop sec pour être travaillé ou les meubles qui ont besoin d’être remis en état.
✓	Hybrid SMT	Cette crème nettoyante peut être utilisée pour hydrater des **démangeaisons de la peau ou le bois**, y compris le bois qui est trop SEC pour le travail du bois ou des meubles qui doivent être restaurés.
✓	CNN	Que la crème tout-en-un peut être utilisée pour hydrater **la peau de démangeaison ou le bois**, y compris Le bois qui est trop sec pour le travail du bois ou les meubles qui doivent être restaurés.
✗*	RNN	Cette crème tout-en-un peut être utilisée pour hydrater **la peau ou le bois**, y compris le bois trop sec pour le travail du bois ou les meubles qui doivent être restaurés.

✓	Attention	Que la crème tout-en-un puisse être utilisée pour hydrater la **peau d' itchy ou le bois**, y compris le bois trop sec pour le travail du bois ou les meubles qui doivent être restaurés.
✗	Google	Cette crème tout-en-un peut être utilisée pour hydrater **la peau ou le bois qui démangent**, y compris le bois trop sec pour le travail du bois ou les meubles qui doivent être restaurés.
✗	DeepL	Cette crème tout-en-un peut être utilisée pour hydrater **la peau ou le bois qui démange**, y compris le bois trop sec pour le travail du bois ou les meubles qui doivent être restaurés.
S15g	Source	I have many pets at home, including **fluffy dogs and geckos** that we keep in different rooms at night so they can have a good night's sleep.
	Ref	J'ai beaucoup d'animaux de compagnie à la maison, y compris des **chiens poilus et des geckos** que je laisse dormir dans différentes pièces la nuit, pour qu'ils aient une bonne nuit de sommeil.
✓	Hybrid SMT	J' ai beaucoup d' animaux de compagnie à la maison, y compris **les chiens duveteuses et les geckos** que nous gardons dans différentes salles de nuit afin qu' ils puissent avoir une bonne nuit de sommeil 's.
✗*	CNN	J' ai beaucoup d' animaux de compagnie à la maison, y compris **les chiens et les geckos** que nous conservons L' hôtel est très bien situé, à proximité du centre ville.
✗*	RNN	J' ai beaucoup d' animaux de compagnie à la maison, y compris **les chiens et les geckos**, que nous conservons dans différentes pièces la nuit afin qu' ils puissent avoir une bonne nuit de sommeil.
✗*	Attention	J' ai beaucoup d' animaux de compagnie à la maison, y compris **des chiens et des geckos** que nous conservons dans différentes chambres la nuit, afin qu' ils puissent avoir une bonne nuit de sommeil.
✓	Google	J'ai beaucoup d'animaux de compagnie à la maison, y compris **des chiens moelleux et des geckos** que nous gardons dans différentes pièces la nuit pour qu'ils puissent passer une bonne nuit.

✓	DeepL	J'ai beaucoup d'animaux domestiques à la maison, notamment **des chiens en peluche et des geckos** que nous gardons dans différentes pièces la nuit pour qu'ils puissent passer une bonne nuit de sommeil.
S15h	Source	I ordered all of my **wood blinds and curtains** online because they were cheaper and I have many windows in my new home that are facing the street.
	Ref	J'ai commandé des **stores en bois et des rideaux** en ligne parce qu'ils coûtaient moins cher et j'ai de nombreuses fenêtres dans ma nouvelle maison qui donnent sur la rue.
✗*	Hybrid SMT	J' ai ordonné à tous mes **stores et rideaux** en ligne parce qu' ils étaient moins chers et j' ai beaucoup de fenêtres dans ma nouvelle maison qui sont confrontés à la rue.
✗*	CNN	J' ai commandé tous mes **stores et rideaux** en ligne parce qu' ils étaient moins chers et j' en ai beaucoup les fenêtres de ma nouvelle maison qui font face à la rue.
✓	RNN	J' ai commandé tous mes **stores de bois et mes rideaux** en ligne parce qu' ils étaient moins chers et que j' ai beaucoup de fenêtres dans ma nouvelle maison qui font face à la rue.
✗	Attention	J' ai commandé tous mes **stores et rideaux de bois en ligne** parce qu' ils étaient moins chers et j' ai beaucoup de fenêtres dans ma nouvelle maison qui sont face à la rue.
✗	Google	J'ai commandé tous mes **stores et rideaux en bois** en ligne car ils étaient moins chers et j'ai de nombreuses fenêtres dans ma nouvelle maison qui donnent sur la rue.
✗	DeepL	J'ai commandé tous mes **stores et rideaux en bois** en ligne parce qu'ils étaient moins chers et que j'ai beaucoup de fenêtres dans ma nouvelle maison qui donnent sur la rue.

Anaphora

"It"

Short sentences		
Without interruption		
S16a	Source	I have a plant pot for my mint, otherwise **it** grows all over the garden.
	Ref	J'ai un pot pour ma menthe, sinon **elle** pousse partout dans le jardin.
✓	Hybrid SMT	J' ai un pot pour ma monnaie, sinon **elle** pousse partout dans le jardin.
✓*	CNN	J' ai un pot de plante pour mon menthe, sinon **elle** pousse dans tout le jardin.
✓*	RNN	J' ai un pot d' usine pour mon menthe, sinon **elle** pousse dans tout le jardin.
✗*	Attention	J' ai un pot de plante pour mon minet, sinon **il** pousse partout dans le jardin.
✗	Google	J'ai un pot de fleurs pour ma menthe, sinon **il** pousse partout dans le jardin.
✓	DeepL	J'ai un pot de plantes pour ma menthe, sinon **elle** pousse partout dans le jardin.
S16b	Source	I wanted the bag with the side pocket because **it** adds a lot of space.
	Ref	Je voulais le sac avec la poche latérale parce qu'**elle** rajoute beaucoup d'espace.
✓	Hybrid SMT	Je voulais le sac avec la poche de côté, car **elle** ajoute beaucoup d' espace.
✗	CNN	J' ai voulu le sac avec la poche latérale parce qu' **il** ajoute beaucoup d' espace.
✗	RNN	J' ai voulu le sac avec la poche latérale parce qu' **il** ajoute beaucoup d' espace.
✗	Attention	Je voulais le sac avec la poche latérale parce qu' **il** ajoute beaucoup d' espace.
✗	Google	Je voulais le sac avec la poche latérale car **il** ajoute beaucoup d'espace.
✗	DeepL	Je voulais le sac avec la pochette latérale parce qu'**il** ajoute beaucoup d'espace.

With interruption		
S16c	Source	He put his brother’s photo in a book after breaking the frame **it** was in.
	Ref	Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequelle **elle** se trouvait.
✗	Hybrid SMT	Il a mis son frère 's photo dans un livre après la rupture du cadre qu' **il** occupait.
✗	CNN	Il a mis la photo de son frère dans un livre après avoir rompu le cadre qu' **il** était.
✗	RNN	Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequel **il** était.
✗	Attention	Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequel **il** se trouvait.
✗	Google	Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequel **il** se trouvait.
✓	DeepL	Il a mis la photo de son frère dans un livre après avoir brisé le cadre dans lequel **elle** se trouvait.
S16d	Source	The recital was tonight but there were issues with the stage so **it** was cancelled.
	Ref	Le récital était ce soir, mais il y avait des problèmes avec la scène alors **il** a été **annulé**.
✓	Hybrid SMT	Le considérant a été ce soir, mais il y a eu des problèmes avec la scène de sorte qu' **il** a été annulé.
✓	CNN	Le considérant était ce soir, mais il y avait des problèmes avec le stade, alors **il** a été annulé.
✓	RNN	Le considérant était ce soir, mais il y avait des problèmes avec l' étape, de sorte qu' **il** a été annulé.
✓	Attention	Le considérant était ce soir, mais il y avait des problèmes avec la scène, de sorte qu' **il** a été annulé.
✓	Google	Le récital était ce soir mais il y avait des problèmes avec la scène donc **il** a été annulé.
✓	DeepL	Le récital était ce soir mais il y a eu des problèmes avec la scène, alors **il** a été annulé.

Long sentences		
Without interruption		
S16e	Source	I use a lot of herbs in my recipes so I have a plant pot for my mint, otherwise **it** grows all over the garden and I can't grow anything else.
	Ref	J'utilise beaucoup de fines herbes dans mes recettes alors j'ai un pot pour ma menthe, sinon **elle** pousse partout dans le jardin et je ne peux rien faire pousser d'autre.
✓	Hybrid SMT	J' utilise beaucoup d' herbes dans mes recettes si j' ai un pot pour ma monnaie, sinon **elle** pousse partout dans le jardin et je can't cultiver autre chose.
✗	CNN	J' utilise beaucoup d' herbes dans mes recettes, j' ai un pot de plante pour mon cuir, sinon **elle** pousse partout dans le jardin et je ne peux rien faire d' autre.
✓*	RNN	J' utilise beaucoup d' herbes dans mes recettes, donc j' ai un pot de plante pour mon menthe, sinon **elle** pousse sur le jardin et je ne peux rien pousser autrement.
✗*	Attention	J' utilise beaucoup d' herbes dans mes recettes, donc j' ai un pot de plante pour mon minet, sinon **il** pousse partout dans le jardin et je ne peux rien faire d' autre.
✓	Google	J'utilise beaucoup d'herbes dans mes recettes donc j'ai un pot de fleurs pour ma menthe, sinon **elle** pousse dans tout le jardin et je ne peux rien faire d'autre.
✓	DeepL	J'utilise beaucoup d'herbes dans mes recettes, donc j'ai un pot de plantes pour ma menthe, sinon **elle** pousse partout dans le jardin et je ne peux rien faire pousser d'autre.
S16f	Source	I wanted the brown leather bag with the white side pocket because **it** adds a lot of space and I can easily reach my phone there.
	Ref	Je voulais le sac de cuir brun avec la poche latérale blanche parce qu'**elle** rajoute beaucoup d'espace et je peux facilement accéder à mon téléphone de là.
✓	Hybrid SMT	Je voulais que le sac en cuir brun avec le côté blanc de poche, car **elle** ajoute beaucoup d' espace et je peux facilement atteindre mon téléphone.
✗	CNN	Je voulais le sac en cuir brun avec la poche blanche, car **il** ajoute beaucoup d' espace et I Vous pouvez facilement rejoindre mon téléphone là-bas.

✗*	RNN	Je voulais que le sac de cuir brun avec la pochette latérale blanche ajoute beaucoup d' espace et je peux facilement atteindre mon téléphone là-bas.
✗	Attention	Je voulais le sac de cuir brun avec la poche latérale blanche parce qu' **il** ajoute beaucoup d' espace et je peux facilement rejoindre mon téléphone là-bas.
✗	Google	Je voulais le sac en cuir marron avec la poche latérale blanche car **il** ajoute beaucoup d'espace et je peux facilement y joindre mon téléphone.
✗	DeepL	Je voulais le sac en cuir marron avec la poche latérale blanche parce qu'**il** ajoute beaucoup d'espace et que je peux facilement y atteindre mon téléphone.
With interruption		
S16g	Source	He put his little brother's graduation photo in a book with all of their childhood photos after accidentally breaking the wooden frame **it** was in.
	Ref	Il a mis la photo de finissant de son petit frère dans un livre avec toutes leurs photos d'enfance après avoir brisé le cadre de bois dans lequel **elle** se trouvait.
✗	Hybrid SMT	Il a mis son petit frère 's photo de graduation dans un livre avec toutes les photos de leur enfance après avoir accidentellement briser le cadre en bois qu' **il** occupait.
✗	CNN	Il a mis la photo de son petit frère dans un livre avec toutes leurs photos d' enfance après l' obtention d' un diplôme de son petit frère. s' enfonce accidentellement dans le cadre en bois qu' **il** était en bois.
✓	RNN	Il a mis la photo de son petit frère dans un livre avec toutes leurs photos d' enfance après avoir rompu accidentellement le cadre en bois qu' **elle** était.
✗	Attention	Il a mis la photo de graduation de son petit frère dans un livre avec toutes ses photos d' enfance après avoir brisé accidentellement le cadre en bois qu' **il** était dans.
✗	Google	Il a mis la photo de remise des diplômes de son petit frère dans un livre avec toutes leurs photos d'enfance après avoir accidentellement brisé le cadre en bois dans lequel **il** se trouvait.
✓	DeepL	Il a mis la photo de fin d'études de son petit frère dans un livre avec toutes leurs photos d'enfance après avoir accidentellement cassé le cadre en bois dans lequel **elle** se trouvait.

S16h	Source	The piano recital was supposed to be tonight in the concert hall, but there were lighting issues with the new stage so **it** was cancelled.
	Ref	Le récital de piano était supposé être dans la salle de concert ce soir, mais il y avait des problèmes avec la nouvelle scène alors **il** a été annulé.
✓	Hybrid SMT	Le récital de piano était censé être ce soir, dans la salle de concert, mais il y avait des questions d' éclairage avec la nouvelle étape de sorte qu' **il** a été annulé.
✓	CNN	Le récital de piano était censé être ce soir dans la salle de concert, mais il y avait des problèmes d' éclairage: la nouvelle étape, c' est qu' **il** a été annulé.
✓	RNN	Le récital de piano était censé être ce soir dans la salle de concert, mais il y avait des problèmes d' éclairage avec la nouvelle scène pour qu' **il** soit annulé.
✓	Attention	Le considérant de piano était censé être ce soir dans la salle de concert, mais il y avait des problèmes d' éclairage avec la nouvelle scène, de sorte qu' **il** a été annulé.
✓	Google	Le récital de piano était censé avoir lieu ce soir dans la salle de concert, mais il y avait des problèmes d'éclairage avec la nouvelle scène, donc **il** a été annulé.
✓	DeepL	Le récital de piano devait avoir lieu ce soir dans la salle de concert, mais il y a eu des problèmes d'éclairage avec la nouvelle scène, donc **il** a été annulé.

"They"

Short sentences		
Cue before pronoun		
S17a	Source	The matriarchs are the rulers of the family and **they** usually make important decisions.
	Ref	Les matriarches sont les dirigeantes de la famille et **elles** prennent normalement les décisions importantes.
✗	Hybrid SMT	Les matriarches sont les gouvernants de la famille et **ils** font généralement des décisions importantes.
✗	CNN	Les matriarches sont les dirigeants de la famille et **ils** prennent habituellement des décisions importantes.
✗	RNN	Les matriarches sont les dirigeants de la famille et **ils** prennent habituellement des décisions importantes.

✗	Attention	Les matriarches sont les dirigeants de la famille et **ils** prennent habituellement des décisions importantes.
✓*	Google	Les matriarches sont les dirigeants de la famille et prennent généralement des décisions importantes.
✓*	DeepL	Les matriarches sont les chefs de famille et prennent généralement les décisions importantes.
S17b	Source	The waitresses came with the bill, but **they** forgot to include some of our drinks.
	Ref	Les serveuses sont venues avec l'addition, mais **elles** ont oublié d'inclure certaines de nos boissons.
✗	Hybrid SMT	Les serveuses sont venus avec le projet de loi, mais **ils** ont oublié d' inclure certains de nos boissons.
✗	CNN	Les serveuses sont venues avec le projet de loi, mais **ils** ont oublié d' inclure certaines de nos boissons.
✓	RNN	Les serveuses sont arrivées avec la facture, mais **elles** ont oublié d' inclure certaines de nos boissons.
✗	Attention	Les serveurs sont arrivés avec le projet de loi, mais **ils** ont oublié d' inclure certaines de nos boissons.
✓	Google	Les serveuses sont venues avec l'addition, mais **elles** ont oublié d'inclure certaines de nos boissons.
✓	DeepL	Les serveuses sont venues avec l'addition, mais **elles** ont oublié d'inclure certaines de nos boissons.
Cue after pronoun		
S17c	Source	The nurses are unhappy **they** got an extra shift after coming back from paternity leave.
	Ref	Les infirmiers sont mécontents qu'**ils** aient reçu un quart de travail supplémentaire après être revenus de leur congé de paternité.
✗	Hybrid SMT	Les infirmières sont mécontents, **ils** ont obtenu un changement supplémentaire au retour du congé de paternité.
✓*	CNN	Les infirmières et infirmiers sont malheureux qu' **ils** aient eu un changement supplémentaire après avoir reculé du congé de paternité.
✗	RNN	Les infirmières ne sont pas satisfaites qu' **elles** ont eu un quart de travail supplémentaire après avoir quitté leur congé de paternité.
✗	Attention	Les infirmières sont malheureuses qu' **elles** aient obtenu un quart de travail supplémentaire après leur retour en congé de paternité.

✗*	Google	Les infirmières sont mécontentes d'avoir eu un quart de travail supplémentaire après leur retour de congé de paternité.
✗*	DeepL	Les infirmières sont mécontentes d'avoir obtenu un poste supplémentaire après leur retour de congé de paternité.
S17d	Source	When those two joined that team, **they** turned the all-male team into a mixed one.
	Ref	Quand ces deux-là **se sont jointes** à l'équipe, **elles** ont fait de l'équipe entièrement masculine une équipe mixte.
✗	Hybrid SMT	Lorsque ces deux ont rejoint cette équipe, **ils** ont transformé l' équipe composée entièrement en une mixte.
✗	CNN	Lorsque ces deux équipes se sont jointes à cette équipe, **ils** ont transformé l' équipe tout-pale en une équipe mixte.
✓	RNN	Lorsque ces deux personnes **se sont jointes** à l' équipe, **elles** ont transformé l' équipe de tous les hommes en une équipe mixte.
✓	Attention	Lorsque ces deux personnes **se sont jointes** à cette équipe, **elles** ont transformé l' équipe entièrement masculine en une équipe mixte.
✗	Google	Lorsque ces deux-là **ont rejoint** cette équipe, **ils** ont transformé l'équipe entièrement masculine en une équipe mixte.
✗	DeepL	Lorsque ces deux-là **ont rejoint** l'équipe, **ils** ont transformé l'équipe exclusivement masculine en une équipe mixte.
Long sentences		
Cue before pronoun		
S17e	Source	Matriarchs are the rulers of the family and **they** usually make important decisions, as property, land, and inheritance are passed down from mother to daughter in their society.
	Ref	Les matriarches sont les dirigeantes de la famille et **elles** prennent normalement les décisions importantes, étant donné que les propriétés, les terrains et les héritages sont transmis de mère en fille dans leur société.
✗	Hybrid SMT	Les matriarches sont les gouvernants de la famille et **ils** font généralement des décisions importantes, comme la propriété, la terre et l' héritage se transmettent de mère en fille au sein de leur société.

✗	CNN	Les matriarches sont les dirigeants de la famille et **ils** prennent habituellement des décisions importantes, comme suit: les biens, les terres et l' hérédité sont transmises de la mère à la fille dans leur société.
✗	RNN	Les matriarches sont les dirigeants de la famille et **ils** prennent habituellement d' importantes décisions, car la propriété, la terre et l' héritage sont transmis de la mère à la fille dans leur société.
✓*	Attention	Les matriarches sont les dirigeants de la famille et prennent habituellement des décisions importantes, car les biens, les terres et l' héritage sont transmis de la mère à la fille dans leur société.
✓	Google	Les matriarches sont les dirigeants de la famille et **elles** prennent généralement des décisions importantes, car la propriété, la terre et l'héritage sont transmis de mère en fille dans leur société.
✓*	DeepL	Les matriarches sont les chefs de famille et prennent généralement des décisions importantes, car les biens, les terres et l'héritage sont transmis de mère en fille dans leur société.
S17f	Source	The waitresses came with the bill for our table but **they** forgot to include some of our drinks, and the desserts that we ordered were not on the check either.
	Ref	Les serveuses sont venues avec l'addition pour notre table, mais **elles** ont oublié d'inclure certaines de nos boissons et les desserts que nous avions commandés n'y figuraient pas non plus.
✗	Hybrid SMT	Les serveuses sont venus avec la facture de notre table, mais **ils** ont oublié d' inclure certains de nos boissons, et les desserts que nous avons ordonné n' étaient pas non plus de contrôle.
✗	CNN	Les serveuses sont venus avec le projet de loi pour notre table, mais **ils** ont oublié d' inclure certaines de nos boissons., et les desserts que nous avons commandés n' étaient pas à l' épreuve.
✗*	RNN	The staff were very friendly and helpful. The room was clean and comfortable.

✗*	Attention	Les serveurs étaient venus avec la facture pour notre table, mais **ils** ont oublié d' inclure certaines de nos boissons, et les desserts que nous avons commandés n' étaient pas non plus sur la vérification.
✓	Google	Les serveuses sont venues avec l'addition de notre table mais **elles** ont oublié d'inclure certaines de nos boissons, et les desserts que nous avons commandés n'étaient pas non plus sur le chèque.
✓	DeepL	Les serveuses sont venues avec la note de notre table, mais **elles** ont oublié d'inclure certaines de nos boissons, et les desserts que nous avons commandés n'étaient pas non plus sur l'addition.
Cue after pronoun		
S17g	Source	The intensive care unit nurses are unhappy and in shock that **they** got an extra shift barely a day after coming back from paternity leave.
	Ref	Les infirmiers de l'unité des soins intensifs sont à la fois mécontents et sous le choc qu'**ils** aient reçu un quart de travail supplémentaire, un jour à peine après être revenus de leur congé de paternité.
✗	Hybrid SMT	Les infirmières de soins intensifs sont mécontents et en état de choc qu' **ils** ont obtenu un changement supplémentaire à peine un jour après son retour de congé de paternité.
✗	CNN	Les infirmières de l' unité de soins intensifs sont malheureusement malheureusement qu' **ils** n' ont plus qu' un quart de travail supplémentaire. jour après le retour du congé de paternité.
✗	RNN	Les infirmières de l' unité de soins intensifs sont mécontentes et en choc qu' **elles** obtiennent un quart de travail supplémentaire à peine après leur retour du congé de paternité.
✗	Attention	Les infirmières de l' unité de soins intensifs sont malheureuses et choquantes de constater qu' **elles** ont subi un déplacement supplémentaire à peine un jour après avoir pris un congé de paternité.
✗*	Google	Les infirmières de l'unité de soins intensifs sont mécontentes et choquées d'avoir eu un quart de travail supplémentaire à peine un jour après leur retour de congé de paternité.

✗*	DeepL	Les infirmières de l'unité de soins intensifs sont malheureuses et choquées d'avoir obtenu un poste supplémentaire à peine un jour après leur retour de congé de paternité.
S17h	Source	When those two joined that software engineering team of four, **they** turned the formerly all-male team into a mixed one that is now one-third female.
	Ref	Quand ces deux-là se sont jointes à l'équipe de génie logiciel de quatre personnes, **elles** ont fait de l'équipe anciennement entièrement masculine une équipe mixte, dont le tiers est composé de femmes.
✗	Hybrid SMT	Lorsque ces deux logiciels ont rejoint que l' équipe d' ingénieurs de quatre, **ils** ont transformé l' ancienne équipe composée entièrement dans une mixte qui est maintenant un tiers des femmes.
✗	CNN	Lorsque ces deux équipes se sont jointes à l' équipe d' ingénierie logicielle de quatre, **ils** ont transformé l' anciennement tout-male l' équipe dans un groupe mixte qui représente maintenant un tiers de femmes.
✗*	RNN	Lorsque ces deux hommes se sont joints à l' équipe de génie logiciel de quatre personnes, **ils** ont transformé l' équipe tout-puissante en une équipe mixte qui est maintenant un tiers de la femme.
✓	Attention	Lorsque ces deux personnes se sont jointes à l' équipe d' ingénierie logicielle de quatre personnes, **elles** ont transformé l' ancienne équipe entièrement masculine en une équipe mixte qui est aujourd' hui un tiers féminin.
✗	Google	Lorsque ces deux-là ont rejoint cette équipe d'ingénierie logicielle de quatre personnes, **ils** ont transformé l'équipe autrefois entièrement masculine en une équipe mixte qui compte désormais un tiers de femmes.
✗	DeepL	Lorsque ces deux-là ont rejoint cette équipe de quatre ingénieurs en informatique, **ils** ont transformé l'équipe autrefois exclusivement masculine en une équipe mixte qui est maintenant composée d'un tiers de femmes.

"These"

Short sentences		
Without interruption		
S18a	Source	Girls usually carry handbags because their pockets are too small, but **these** are very spacious.
	Ref	Les filles portent souvent des sacs à main parce que leurs poches sont trop petites, **celles-ci** par contre sont très spacieuses.
✗	Hybrid SMT	Les filles sont généralement porteurs de sacs à main parce que leurs poches sont trop petites, mais **elles** sont très spacieuses.
✗	CNN	Les filles portent habituellement des sacs à main parce que leurs poches sont trop petites, mais **elles** sont très petites. spacieux.
✗	RNN	Les filles transportent habituellement des sacs à main parce que leurs poches sont trop petites, mais **elles** sont très spacieuses.
✗	Attention	Les filles portent habituellement des sacs à main parce que leurs poches sont trop petites, mais **elles** sont très spacieuses.
✓	Google	Les filles portent généralement des sacs à main parce que leurs poches sont trop petites, mais **celles-ci** sont très spacieuses.
✗	DeepL	Les filles portent généralement des sacs à main parce que leurs poches sont trop petites, mais **ceux-ci** sont très spacieux.
S18b	Source	Most jewellers sell lots of engagement rings, and **these** are always over-priced.
	Ref	La plupart des bijoutiers vendent beaucoup de bagues de fiançailles et **celles-ci** sont toujours hors de prix.
✓	Hybrid SMT	La plupart des bijoutiers vendent beaucoup de bagues et **celles-ci** sont toujours trop chers.
✓*	CNN	La plupart des bijoutiers vendent beaucoup d' anneaux d' engagement, **qui** sont toujours surévalués.
✓*	RNN	La plupart des bijoutiers vendent beaucoup de cercles d' engagement, et **ceux-ci** sont toujours sur-évalués.
✗	Attention	La plupart des bijoux vendent beaucoup d' anneaux d' engagement, et **ils** sont toujours surévalués.
✓	Google	La plupart des bijoutiers vendent beaucoup de bagues de fiançailles, et **celles-ci** sont toujours trop chères.

✓	DeepL	La plupart des bijoutiers vendent beaucoup de bagues de fiançailles, et **celles-ci** sont toujours hors de prix.
With interruption		
S18c	Source	Most cottage owners are residents of nearby cities, but **these** came from England.
	Ref	La plupart des propriétaires de chalet sont des résidents de villes avoisinantes, **ceux-ci,** par contre, viennent d'Angleterre.
✓	Hybrid SMT	La plupart des propriétaires de chalets sont des résidents des villes voisines, mais **ceux-ci** provenaient de l' Angleterre.
✗	CNN	La plupart des propriétaires de chalets sont des résidents des villes voisines, mais **ils** proviennent de l' Angleterre.
✓	RNN	La plupart des propriétaires de chalets sont des résidents des villes avoisinantes, mais **ceux-ci** viennent d' Angleterre.
✗	Attention	La plupart des propriétaires de chalets sont des résidents des villes voisines, mais **ils** viennent d' Angleterre.
✓	Google	La plupart des propriétaires de chalets sont des résidents des villes voisines, mais **ceux-ci** venaient d'Angleterre.
✓	DeepL	La plupart des propriétaires de chalets sont des résidents des villes voisines, mais **ceux-ci** viennent d'Angleterre.
S18d	Source	I thought I owned the ugliest shoes in three counties, but **these** are even uglier.
	Ref	Je pensais posséder les chaussures les plus laides de trois comtés, mais **celles-ci** sont encore plus laides.
✓	Hybrid SMT	Je pensais que je possédais les chaussures les plus laids dans trois comtés, mais **celles-ci** sont encore plus moche.
✗*	CNN	J' ai pensé que j' étais propriétaire des chaussures uglies dans trois comtés, mais il s' agit même d' un bout à l' autre. Il s' agit d' un problème de santé.
✓	RNN	J' ai pensé que j' avais les plus beaux chaussures dans trois comtés, mais **ces chaussures** sont même ugelles.
✗	Attention	J' ai pensé posséder les chaussures les plus uglies dans trois comtés, mais **elles** sont même plus aberrantes.
✓	Google	Je pensais posséder les chaussures les plus laides de trois pays, mais **celles-ci** sont encore plus laides.
✓	DeepL	Je pensais que je possédais les chaussures les plus laides de trois comtés, mais **celles-ci** sont encore plus laides.

Long sentences		
Without interruption		
S18e	Source	Most of the girls I know usually carry big handbags that can fit their wallet because their pants pockets are too small, but **these** are very spacious.
	Ref	La plupart des filles que je connais portent souvent de grands sacs à main capables de contenir leur porte-feuille parce que leurs poches sont trop petites, **celles-ci** par contre sont très spacieuses.
✗	Hybrid SMT	La plupart des filles, je sais habituellement transporter de gros sacs qui peuvent tenir leur porte-monnaie, car leurs poches de pantalon sont trop petites, mais **elles** sont très spacieuses.
✗	CNN	La plupart des filles que je connais portent habituellement de gros sacs à main qui peuvent s' ajuster à leur portefeuille parce que leur p Les poches sont trop petites, mais **elles** sont très spacieuses.
✗	RNN	La plupart des filles que je connais portent habituellement de gros sacs à main qui peuvent convenir à leur portefeuille parce que leurs poches de pantalon sont trop petites, mais **elles** sont très spacieuses.
✗	Attention	La plupart des filles que je connais portent habituellement de gros sacs à main qui peuvent s' adapter à leur porte-monnaie parce que leurs poches de pantalons sont trop petites, mais **elles** sont très spacieuses.
✓	Google	La plupart des filles que je connais portent généralement de gros sacs à main qui peuvent rentrer dans leur portefeuille parce que les poches de leurs pantalons sont trop petites, mais **celles-ci** sont très spacieuses.
✗	DeepL	La plupart des filles que je connais portent généralement de grands sacs à main qui peuvent tenir dans leur portefeuille parce que les poches de leurs pantalons sont trop petites, mais **ceux-ci** sont très spacieux.
S18f	Source	Most jewellers in big cities such as Toronto sell lots of nice, shiny, beautifully presented wedding bands and engagement rings, and **these** are always over-priced.
	Ref	La plupart des bijoutiers dans les grandes villes comme Toronto vendent beaucoup d'alliances de mariage et de bagues de fiançailles jolies, brillantes et magnifiquement présentées, et **celles-ci** sont toujours hors de prix.

✗	Hybrid SMT	La plupart des bijoutiers dans les grandes villes comme Toronto vendent des lots de Nice, brillant, magnifiquement présenté des bandes de mariage et de fiançailles, et **ceux-ci** sont toujours trop chers.
✗*	CNN	La plupart des bijoutiers dans les grandes villes, comme Toronto, vendent beaucoup de beaux et brillants L' hôtel est très bien situé, à proximité de l' aéroport.
✓*	RNN	La plupart des bijoutiers des grandes villes comme Toronto vendent beaucoup de belles bandes de mariage et d' anneaux d' engagement magnifiques et joliment présentés, et **ceux-ci** sont toujours sur-évalués.
✗	Attention	La plupart des joailliers des grandes villes comme Toronto vendent beaucoup de belles bandes de mariage et d' anneaux d' engagement bien présentés, et **ils** sont toujours surélevés.
✓	Google	La plupart des bijoutiers des grandes villes comme Toronto vendent beaucoup d'alliances et de bagues de fiançailles jolies, brillantes et joliment présentées, et **celles-ci** sont toujours trop chères.
✓	DeepL	La plupart des bijoutiers des grandes villes, comme Toronto, vendent beaucoup de belles bagues de mariage et de fiançailles, brillantes et bien présentées, et **celles-ci** sont toujours hors de prix.
With interruption		
S18g	Source	Most cottage owners are residents of nearby cities who want a quiet place to relax that is also not too far of a drive, but **these** came from England.
	Ref	La plupart des propriétaires de chalet sont des résidents de villes avoisinantes qui veulent un coin tranquille où relaxer qui ne soit pas à une trop grande distance de route, **ceux-ci**, par contre, viennent d'Angleterre.
✓	Hybrid SMT	La plupart des propriétaires de chalets sont des résidents des villes voisines qui veulent un endroit tranquille pour se détendre qui n' est pas non plus trop loin d' une voiture, mais **ceux-ci** provenaient de l' Angleterre.
✓	CNN	La plupart des propriétaires de chalets sont des résidents des villes voisines qui veulent un endroit tranquille pour se détendre, ce qui n' est pas trop loin. d' un disque, mais **ceux-ci** proviennent de l' Angleterre.

✓	RNN	La plupart des propriétaires de chalets sont des résidents des villes voisines qui veulent un endroit tranquille pour se détendre qui n' est pas trop loin d' une voiture, mais **ceux-ci** viennent d' Angleterre.
✗*	Attention	La plupart des propriétaires de chalets sont des résidents des villes voisines qui veulent un endroit tranquille pour se détendre, ce qui n' est pas trop loin d' une voiture, mais viennent d' Angleterre.
✗	Google	La plupart des propriétaires de chalets sont des résidents de villes voisines qui veulent un endroit calme pour se détendre, pas trop loin en voiture, mais **ils** viennent d'Angleterre.
✓	DeepL	La plupart des propriétaires de chalets sont des résidents de villes voisines qui veulent un endroit tranquille pour se détendre, qui ne soit pas trop éloigné de la route, mais **ceux-ci** viennent d'Angleterre.
S18h	Source	I thought I owned the ugliest shoes in three counties, with the weird lace colour, ridiculous platform, and tacky pattern, but **these** are even uglier.
	Ref	Je pensais posséder les chaussures les plus laides de trois comtés, ornées d'une couleur de lacets étrange, d'une plateforme ridicule et d'un motif ringard, mais **celles-ci** sont encore plus laides.
✓	Hybrid SMT	Je pensais que je possédais les chaussures les plus laids dans trois comtés, avec la couleur bizarre de dentelle, plate-forme ridicule, et le modèle kitsch, mais **celles-ci** sont encore plus moche.
✗	CNN	J' ai pensé que j' étais propriétaire des chaussures uglies dans trois comtés, avec le barrage la lace, la plate-forme ridicule et le schéma de traîne, mais Il s' agit même d' une horreur.
✗*	RNN	J' ai pensé que j' avais les chaussures les moins chères en trois comtés, avec la couleur des dentelles bizarres, la plate-forme ridicule, et les motifs.
✗*	Attention	J' ai pensé posséder les chaussures les plus uglies dans trois comtés, avec la couleur de lace, une plate-forme ridicule, et des motifs aberrants, mais même aberrants.
✓	Google	Je pensais posséder les chaussures les plus laides de trois comtés, avec une couleur de dentelle étrange, une plate-forme ridicule et un motif collant, mais **celles-ci** sont encore plus laides.

✓	DeepL	Je pensais que je possédais les chaussures les plus laides de trois comtés, avec leur étrange couleur de dentelle, leur plateforme ridicule et leur motif collant, mais **celles-ci** sont encore plus laides.