Examining the Quality of French Machine Translation

For this assignment, I will be analysing the French article «Traduction automatique : les réseaux de neurones à l'essai», which can be found here. It is important to note that I am not fluent in French and that my exposure with the language comes primarily from taking 2 semesters of undergraduate-level French coursework.

I was impressed with the quality of French-to-English translation demonstrated by Google Translate. The key problems I noticed with Google Translate are in verb translation: French doesn't differentiate between verbs of the form 'get(s)' and 'are getting', so information on which English form to use in the translation is not encoded in the French verb. More problems can be found with articles, specifically with deciding whether to insert or delete a noun's determiner in the translation, since French nouns largely require a determiner, which is not the case for English nouns. Another problem is that of antecedent ambiguity in the translation, which is especially evident in pronoun translation. This is due partly to the permutations of singular/plural, 1st/2nd/3rd person, and noun gender not aligning in French and English. An example of this is the French *on*, their 3rd.SG.NEUT pronoun of which English has no equivalent. This sort of difference (also seen in articles) contributes to translation ambiguity since information on plurality and noun genders, which demonstrates alignment between words/phrases and their antecedents, is lost in translation. A summary of these problems can be found below.

Verb Translation Errors

Google's translation of ***les chercheurs veulent tester*** *une méthode plus douce* is '**the researchers want try** a gentler method'. 'want try' is grammatical, and a more accurate translation is 'are wanting to try'. This difference is easy to overlook since French doesn't differentiate between the forms for 'want(s)' and 'wanting'.

Similarly, Google's translation of ***Obtenir*** *le meilleur modèle* is **'Get** the best model', but the translation should be 'Getting', especially since this is sub-heading. Again, note that French doesn't differentiate between the verb forms 'get' and 'getting'.

Article Translation Errors

Google's translation of *phrases de médecine ou d'informatique, de tourisme,* ***d'actualités****, de conférences TED* was 'three million sentences on medicine or IT, tourism, **news**, TED talks'. French grammar requires an article before nouns while English grammar doesn't, so these articles are often eliminated in translation. The literal translation of *d'actualities* is 'the news', and the MT of 'news' is an example of unlicensed article deletion. This is one of the few instances the French determiner is kept in the translation since the elimination of the determiner before 'news' is rare.

Antecedent Ambiguity

Google translated *Puis l'« élève » subit une trentaine de tests constitués de 500 phrases à traduire,* ***n'appartenant évidemment pas au corpus d'apprentissage*** as 'Then the "student" undergoes about thirty tests consisting of 500 sentences to be translated, **obviously not belonging to the learning corpus'**. There is ambiguity in the MT translation on what 'obviously not belonging to the learning corpus' refers to. Prior context suggests it refers to the 500 sentences mentioned previously (prior sentence: 'For ten hours, they "ingest" three million sentences on medicine or IT, tourism, **the news**, TED conferences or film subtitles (there are around fifty themes), depending on the choice made by a human'). Switching the ordering of the phrases makes the embedded nature of the clauses clearer: 'Then the "student" undergoes about thirty tests consisting of 500 sentences, **obviously not from the learning corpus**, to be translated.'

Pronoun Antecedent Ambiguity

Google's translation of *C'était une dizaine de jours après le début d'une expérience originale menée dans **son** centre de recherche en traduction automatique* is 'It was about ten days after the start of an original experiment carried out in **its** machine translation research center'. In the translation, it is unclear what 'its' refers to, but from the context, "son" likely refers to the company Systran (previously mentioned in the article), since we know that's where the experiment was conducted. In French, "son" is the 3rd person singular pronoun (gender neutral) so we can use this information to confirm that this pronoun is referring to Systran. To show this, the pronoun 'their' should have been used.

Google's translation of *les professeurs savent quel corpus l'a nourri, par exemple un quart de dialogues, un quart de juridique, **autant d'actualités que de médical*** is **'**the professors know what corpus has fed it, for example a quarter of dialogues, a quarter of legal, **as much news as of medical**', which doesn't make sense in English. It is ambiguous what 'as much news as of medical' means since 'as of medical' doesn't mean anything in English. A better translation is 'as much news as that of medical' since it represents the meaning of *'que'* and adds the missing pronoun which clears the ambiguity.

Pronoun Translation: *on*

Google translated *si **on spécialise** tout de suite l'apprentissage* as '*if **you specialize** the learning right away'*. Interestingly, "*on*" is French's 3.SG.NEUT pronoun. English doesn't have pronouns of that sort, but 'one' would be a better translation than 'you', which is second person.

Google translated *« L'un des problèmes est que si **on** spécialise tout de suite l'apprentissage, le réseau de neurones devient "bête* as '*"One of the problems is that if **we** immediately specializes the learning, the neural network becomes "dumb"'*. This has the same problem with translating "on" except MT translated it as 'we' this time. A better translation would be 'someone' since 'we' is 1.PL.NEUT rather than 3.SG.NEUT .

SOURCE SENTENCE	MACHINE TRANSLATION	CORRECTION	ASSESSMENT
Traduction automatique : les réseaux de neurones à l'essai	Machine translation: neural networks on trial	Machine Translation: Neural Networks on Trial	(Title) The only thing I'd change is capitalizing each word since this is the title of the article.
Eduquer **en même temps** plusieurs	Educating several programs on different	**Simultaneously** educating several	(1) Word choice differences: "en même

programmes sur différents thèmes, leur faire passer des examens, puis garder **les meilleurs** pour l’étape suivante...	themes **at the same time**, giving them exams, then keeping **the best ones** for the next stage... this so-called "soft" method is at the heart of an original experience.	programs on different themes, giving them exams, then keeping **the best** for the next stage…	temps” was translated as ‘at the same time’ but I chose ‘simultaneously’. (2) Pronoun insertion: Google translated “ les meilleurs” as ‘best ones’ while I simply did ‘best’.
cette méthode dite « **douce** » est au coeur d’une expérience originale.	This so-called "**soft**" method is at the heart of an original experience.	This so-called “**gentle**” method is at the heart of an original experience.	(1) I translated “douce” as gentle instead of ‘soft’.
« Tout est en route, nous avons une petite classe avec de bons et de mauvais élèves sur lesquels nous faisons de la sélection », résumait Jean Senellart, directeur technique de Systran, **le 6 octobre**.	*"Everything is on the way, we have a small class with good and bad students from whom we are selecting"* , summarized Jean Senellart, technical director of Systran, **on October 6**.	*“Everything is on the way, we have a small class with good and bad students from whom we are selecting”, summarized Jean Senellart, technical director of Systran, **on October 6th**.*	(1) Formatting difference for dates: I would translate “le 6 octobre” as ‘October 6th’ rather than ‘October 6’
C’était une dizaine de jours après le début d’une expérience originale menée dans **son** centre de recherche en traduction automatique.	It was about ten days after the start of an original experiment carried out in **its** machine translation research center.	It was about ten days after the start of an original experiment conducted in **their** machine translation research center.	(1) Pronoun ambiguity: Google translated “son” as ‘its’ but this doesn’t make sense in English and is ambiguous. (2) Pronoun antecedent ambiguity: there is ambiguity in the translation in what this pronoun is referring to. From the context, “son” likely refers to Systran since we know that’s where the experiment was conducted. In French, “son” is the 3rd person singular pronoun (gender neutral) so we can also use this information to determine that this pronoun is referring to Systran. (3) Pronoun antecedent ambiguity:

			if the sentence is referring to Systran, the better English translation would be ‘their’ since ‘its’ doesn’t make sense here English.
Au lieu de faire ingurgiter plus de dix millions de phrases pendant des semaines à un algorithme de type réseau de neurones pour lui apprendre à traduire l’anglais vers le français, comme tout le monde le fait, y compris Systran depuis 2016, les chercheurs **veulent tester** une méthode plus douce.	Instead of having a neural network-like algorithm ingest more than ten million sentences for weeks to teach it to translate English to French, as everyone else has been doing, including Systran since 2016, the researchers **want try** a gentler method.	Instead of having a neural network algorithm ingest over ten million phrases for weeks to teach it to translate English to French like everyone else has been doing, including Systran since 2016, the researchers **are wanting to try** a gentler method.	(1) Missing word: “les chercheurs veulent tester” was translated ‘want try’ in MT but ‘want to try’ is grammatical. (2) Verb tense: *les chercheurs veulent tester* should have been ‘the researchers are wanting to try’ instead of ‘the researches want try’. ‘want try’ is grammatical. This difference is easy to overlook since French doesn’t differentiate between the forms for ‘want(s)’ and ‘are wanting’. (3): Word choice: MT translated “à un algorithme de type réseau de neurones “ as ‘neural-tetwork-like algorithm’, which doesn’t appear in the literature as much as ‘neural-network algorithm’. (4): Punctuation: the comma after ‘French’ should be removed since it doesn’t make sense in English.
Il s’agit d’éduquer en même temps plusieurs programmes sur différents thèmes, leur faire passer des examens, puis garder	**It is a question of educating several programs at the same time on different themes,** making them pass exams, then	**It is about educating several programs on different themes at the same time,** making them pass exams, then keeping **the best** for the	(1) Unlicensed word insertion: MT translated “Il s’agit d’éduquer “ as ‘it is a question of educating’ even though no word for ‘question’

les meilleurs pour l’étape suivante, identique à **la précédente**, mais avec des phrases de nature différente, et ainsi de suite.	keeping **the best ones** for the next stage, identical to **the previous one**, but with sentences of a different nature, and so on.	next stage, identical to **the former**, but with sentences of a different nature, and so on.	appeared in this sentence. (2) Phrase order: while ‘educating several programs at the same time on different themes’ is what happens literally in the French, this order/bracketing doesn’t make sense in English. (3) Unlicensed pronoun insertion: MT chose ‘best ones’ for “les meilleurs”; ‘ones’ was added by MT. Similar thing with ‘previous one’ for “la precedente”; there is no ‘ones’ in the French and it is not necessary for the English translation to make sense.
Obtenir le meilleur modèle	Get the best model	Getting the best model.	(Subheading) (1) Verb form error: “Obtenir” was translated as ‘get’ by MT but should be ‘getting’. Using the former does not make sense.
Plus concrètement, **au départ**, cinq réseaux de neurones**,** aux 400 millions de paramètres chacun, **tirés aléatoirement**, participent à l’expérience.	More concretely, **at the start**, five neural networks**,** with 400 million parameters each, **drawn randomly**, take part in the experiment.	More concretely, **at the beginning**, five neural networks with 400 million parameters each, **randomly drawn**, take part in the experiment.	(1) Word choice: “au depart” was translated as ‘at the start’ by MT but ‘at the beginning’ makes more sense in English. (2) Punctuation The comma after ‘neural. networks’ was copied from the French but doesn’t make sense in the English. (3): Word choice: ‘drawn randomly’ < ‘randomly drawn’ in English.
Pendant dix heures, ils « ingurgitent » trois	For ten hours, they "ingest" three million	For ten hours, they “ingest” three million	(1) Unlicensed article deletion; “d’actualites”

millions de phrases de médecine ou d’informatique, de tourisme, **d’actualités**, de conférences TED ou de sous-titres de films (il y a une cinquantaine de thèmes), selon le choix fait par un humain.	sentences on medicine or IT, tourism, **news**, TED talks or film subtitles (there are around fifty themes), depending on the choice made**.** by a human.	sentences on medicine or IT, tourism, **the news**, TED conferences or film subtitles (there are around fifty themes), depending on the choice made by a human.	was translated as ‘news’ but in English it should be ‘the news’ (one of the few instances the French determiner is kept). (2) Punctuation: There is an incorrect period added by MT after ‘made’ which doesn’t appear in the French.
Puis l’« élève » subit une trentaine de tests constitués de 500 phrases à traduire, **n’appartenant évidemment pas au corpus d’apprentissage**.	Then the “student” undergoes about thirty tests consisting of 500 sentences to be translated, **obviously not belonging to the learning corpus**.	Then the “student” undergoes about thirty tests consisting of 500 sentences, **obviously not from the learning corpus**, to be translated.	(1) Antecedent ambiguity/phrase ordering: there is ambiguity in the MT translation on what ‘obviously not belonging to the learning corpus’ refers to. Context suggests it refers to the 500 sentences so I moved that clause after ‘sentences’ to make that clear. (2) Word choice: ‘not belonging to the learning corpus’ is also less clear/natural as ‘not from the learning corpus’.
A chaque « époque » – comme ils baptisent la **période apprentissage/test** –, cinq ou six modèles sont gardés pour la phase suivante.	At each “epoch” – as they call the **learning/testing period** – five or six models are kept for the next phase.	At each “epoch” - as they call the **learning/test period -** five or six models are kept for the next phase.	(1) Word choice: The French had “a période apprentissage/test –” which MT translated as ‘learning-testing period’ but ‘test-period’ would be a more accurate translation.
Ils constituent une nouvelle génération qui subira le même sort que ses parents.	**They are** a new generation that will suffer the same fate as their parents.	**They constitute** a new generation that will suffer the same fate as their parents.	(1) Word choice; ‘they constitute’ is a more accurate translation of “ils constituent” than MT’s ‘they are’ which is ambiguous.
« L’un des problèmes est que si ***on spécialise***	*"One of the problems is that if* ***you specialize***	*“One of the problems is that if* ***one specializes***	(1) Pronoun error: “on specialize” was

tout de suite l’apprentissage, le réseau de neurones devient “bête” et peut bloquer sur des traductions simples »	*the learning right away, the neural network becomes 'dumb' and can get stuck on simple translations"*	*the learning right away, the neural network becomes “dumb” and can get stuck on simple translations.”*	translated as ‘you specialize’ by MT but “on” in French is a 3rd person pronoun of neutral gender, not 2nd person. English doesn’t have one of that sort but ‘one’ is the closest you can get.
Une sorte d’écran radar **permet** de surveiller la classe à chaque étape.	A sort of radar screen **makes it possible to** monitor the class at each stage.	A sort of screen radar **allows** class monitoring at each stage.	(1) Verb voicing: MT translated this “permet de surveiller la classe” in passive voice ‘makes it possible to monitor the class’ instead of active ‘allows class monitoring’.
Les élèves sont des ronds, avec un matricule à sept lettres et chiffres, sous lesquels leurs notes apparaissent pour les divers tests.	The students are circles, with a number of seven letters and numbers, under which their marks appear for the various tests.	The students are circles, with a number of seven letters and numbers, under which their marks appear for the various tests.	No problems
En cliquant sur un rond, les professeurs savent quel corpus l’a nourri, par exemple un quart de dialogues, un quart de juridique, **autant d’actualités que de médical**…	By clicking on a circle, the professors know what corpus has fed it, for example a quarter of dialogues, a quarter of legal, **as much news as of medical**...	By clicking on a circle, the professors know what corpus fed it, for example, a quarter of dialogue, a quarter of legal, **as much news as that of medical**.	(1) Pronoun ambiguity: The MT translation for “autant d’actualies que de medical” (’as much news as of medical’) doesn’t make sense in English. ‘as much news as that of medical’ clears the ambiguity.
Des liens entre les ronds permettent de repérer leur ascendance.	Links between the circles make it possible to identify their ancestry.	Links between the circles make it possible to identify their ancestry.	No problems
« L’un des problèmes est que si ***on*** *spécialise tout de suite l’apprentissage, le réseau de neurones devient “bête” et* ***peut bloquer*** *sur des traductions simples »*,	*"One of the problems is that if* ***we*** *immediately specialize the learning, the neural network becomes 'dumb' and* ***can block*** *on simple translations",* notes Jean Senellart, who also	*“One of the problems is that if* ***someone*** *immediately specializes the learning, the neural network becomes “dumb” and* ***can get stuck on*** *simple translations”, notes*	(1) Pronoun error: Same problem translating “on” except MT translated it as ‘we’ this time. (2) Word choice: MT translated “peut bloquer” as ‘can block’ while in the

note Jean Senellart, qui indique aussi que *« 350 phrases sur 13 millions peuvent changer du tout au tout un comportement »*.	indicates that *"350 sentences out of 13 million can completely change a behavior .*	*Jean Senellart, who also indicates that “350 sentences on the 13 million can completely change a behavior”.*	previous appearance of this word it was ‘can get stuck’ which is a better English translation.
Il vous reste 32.92% de cet article à lire.	**You have** 32.92% of this article left to read.	**You still have** 32.92% of this article left to read.	(1) Unlicensed word deletion: MT eliminated ‘still’ in “il vous reste”. It should be ‘you still have’ instead of ‘you have’.
La suite est réservée aux abonnés.	The following is for subscribers **only**.	The following is reserved for subscribers.	(1) Unlicensed word insertion: MT added ‘only’ while translating “aux abonnes”; it should be ‘for subscribers’ not ‘for subscribers only’.