CS124[1] Final Project Report

I. Introduction: The French Language

French and English are two very closely related languages, sharing exact meanings between many words and phrases. This means that we could obtain a reasonably fluent result with lexical translation combined with several rules to correct for structural differences between the two languages. Additionally, the fact that two out of three of our group members have some fluency in French means that we were able to easily identify grammatical patterns that required adjustment for fluent translation. For example, we were immediately able to identify several structural differences between French and English:

1. French uses **gendered nouns** while English does not. For instance, “Il pleut” literally translates to “he rains,” but in English we would simply say “it’s raining.” “Il mange” translates to “he eats,” which is what we would say in English. The fact that French pronouns don’t correspond directly to English pronouns makes translation difficult.
2. **Adjective ordering** in French is usually “adjective noun” but sometimes “noun adjective” in the case of a few specific adjectives; e.g. “chat rouge” and “bon enfant” (literally “cat red” and “good child”). In contrast, adjective ordering in English is always “adjective noun.”
3. **SVO order**, which is “SVO for nouns, but SOV when the object is a pronoun and VSO for questions” (http://en.wikipedia.org/wiki/Word_order), e.g. “Nino aime Amelie,” (Nino loves Amelia) “Je t’aime,” (I love you) “Aimes-tu moi? (Do you love me?)” In contrast, English is generally SVO. Additionally, some indirect objects in English are direct in French and vice-versa. For example, in “Je lui parle,” “lui” is a direct object, while in its translation, “I talk to him,” the word “him” is a direct object.
4. The **present participle** is used differently in French and English. In English, we usually use the present participle to indicate an action that is currently happening, and the simple present to indicate actions that happen habitually: “He is eating cereal” and “He eats cereal every day.” French would use the simple present tense, “Il mange des céréales” in both cases.
5. **Possessives**: French does not use apostrophe-”s” possessives, but instead uses “de.” Although a fluent translation of “le maison de Pierre” is “the house of John,” a better translation would be “John’s house.”
6. **Prepositions**: There is not a one-to-one correspondence between the uses of French and English prepositions. For example, “de” can either mean “from,” “of,” or “about.”
7. **Negatives:** French negative phrases have two negations in them. For example, the phrase “he does **not** eat” in English only has one negative word: “not.” However, the French translation, “Il **ne** mange **pas**” has two negations, “ne” and “pas.”

II. Corpus selection

The Europarl corpus [http://www.statmt.org/europarl/] was chosen because of its relatively simple and correct French grammar and because of the absence of specialized technical or academic jargon such as that found in Wikipedia. Additionally, it provides standard English translations, which can be very useful. We randomly selected 15 sentences and their English translations out of the Europarl corpus of 2007723 sentences. We then randomly chose 10 sentences for the dev set and 5 for the test set. The specific division of the corpus can be found in the “Corpus and Output” section of this write up below.

III. Corpus and Output

Dev set corpus and output (all sentences from the Europarl corpus)

1. French: Monsieur le Président, merci beaucoup de votre compréhension.
Human translation (from Europarl): Mr President, thank you very much for your understanding.

[1] Team Á-1-Ω (alpha-one-omega)

Google Translate: Mr. Speaker, thank you very much for your understanding.
Our machine translation: Mr. presidents thank you many of your comprehension.

2. French: En réalité, tout un système d'organes juridiques et constitutionnels, dont un médiateur, a été mis sur pied en Pologne pour protéger cette liberté et pour garantir le respect de la législation européenne.
Human translation: As a matter of fact, a whole system of judicial and constitutional bodies, including an ombudsman, has been set up in Poland to protect this freedom and to ensure compliance with European legislation.
Google Translate: As a matter of fact, a whole system of judicial and constitutional bodies, including an ombudsman, has been set up in Poland to protect this freedom and to ensure compliance with European legislation.
Our machine translation: In fact all a system legal body and constitutional whose a mediator ha summer placed on foot in Poland for protect this freedom and for ensure the respects of the European legislation.

3. French: Parce que la grand-mère allemande est diabétique, tous les Européens doivent-ils cesser de manger du sucre?
Human translation: Must all Europeans give up sugar just because its German grandmother is diabetic?
Google Translate: Because German grandmother is diabetic, all Europeans have to stop eating sugar?
Our machine translation: Because the German grandma is diabetic all European the must they stop of eats of sugar?

4. French: Je demande donc une fois de plus aux députés de ce Parlement, lorsqu'ils voteront par appel nominal demain, de se prononcer en faveur du droit de l'Assemblée à régir ses propres affaires.
Human translation: I therefore once more appeal today to the Members of this House, when they vote by roll-call tomorrow, to cast a vote for this House's right to govern its own affairs.
Google Translate: So I ask once more members of this Parliament, when voting by roll call tomorrow to vote in favor of the right of the Assembly to govern its own affairs.
Our machine translation: I demand therefore more time to deputys of this parliament when they vote by nominal call tomorrow of its pronounces in favor of right of the assembly to govern its own businesss.

5. French: Je propose que l'on vote sur la version espagnole du texte, en supprimant «de la Presidencia británica».
Human translation: I propose that we vote on the Spanish version of the text, doing away with "of the UK Presidency' .
Google Translate: I propose that we vote on the Spanish version of the text, by deleting "the Presidencia británica."
Our machine translation: I move that one vote on the Spanish version of text in removing " of the Presidencia británica ".

6. French: Nous devons clairement signaler au gouvernement bulgare que le Parlement n'envisagerait même pas de voter un traité d'adhésion si, par exemple, l'article 157 du code pénal, qui établit une discrimination inacceptable pour les citoyens et citoyennes homosexuels de Bulgarie, devait se maintenir.
Human translation: We must say clearly to the Bulgarian Government that this Parliament would not even consider voting for an accession agreement while, for example, Article 157 of the Penal Code remains in force, which establishes unacceptable discrimination against homosexual citizens in Bulgaria.
Google Translate: We must clearly inform the Bulgarian Government and Parliament would not even consider voting an accession treaty if, for example, Article 157 of the Penal Code, which establishes an unacceptable discrimination for gay citizens of Bulgaria, was continue.
Our machine translation: We must clearly report of Bulgarian government that parliament would consider same not of votes a treaty of membership if by example the article 157 of penal code which establish a unacceptable discrimination for the citizens and citizen homosexual of Bulgarian wa its maintain.

7. French: Par ailleurs, au cours des deux prochaines décennies, les augmentations les plus fortes de ces rejets se produiront dans les pays en voie de développement.

Human translation: What is more, the highest increases in emissions during the next two decades will take place in the developing countries.
Google Translate: In addition, over the next two decades, the strongest of these discharges increases will occur in developing countries.
Our machine translation: To somewhere else of course of two next decades the increases the more strong of this releases its produces in the countrys in way of development.

8. French: Je pense qu'il est difficile de surestimer l'importance des marchés de capitaux.
Human translation: It is difficult to overestimate the importance of these capital markets.
Google Translate: I think it is difficult to overestimate the importance of capital markets.
Our machine translation: I think that they is difficult of overestimate the importance of market of capital.

9. French: Je considère que le système de réseaux est déterminant pour l'échange d'informations et de connaissances et pour la réalisation de projets communs entre les diverses zones de l'Union ; je rappelle par ailleurs que pour pouvoir être réalisé pleinement, il doit exploiter, entre autres, tous les instruments qui existent déjà, comme les carrefours.
Human translation: I believe the networking system to be vital for the exchange of information and skills and for the implementation of joint projects between different areas of the Union and I repeat that in order to be fully successful it must exploit inter alia measures already in existence, such as the Carrefours.
Google Translate: I consider that the system of networks is crucial for the exchange of information and knowledge and for the realization of joint projects between the various areas of the Union, I also recalled that in order to be fully realized, it must operate among others, all instruments that already exist, such as intersections.
Our machine translation: I considered that systems of network is determining for the exchange of information and of knowledge and for the achieve of project join between the various areas of the Union, I recall by somewhere else that for power be achieve fully they must exploit between other all instruments which exist already a crossroads.

10. French: La situation en Tchétchénie est mauvaise.
Human translation: The situation in Chechnya is bad.
Google Translate: The situation in Chechnya is bad.
Our machine translation: The situation in Chechnya is bad.

Test set corpus and output **(all sentences from the Europarl corpus)**

1. French: Ce n'est qu'une partie de la contribution de l'UE, et cette contribution doit être en cohérence avec la réaction internationale coordonnée..
Human translation (from Europarl): We need to deliver this as part of a coherent EU contribution within a coordinated international response.
Google Translate: This is only part of the EU contribution, and this contribution must be consistent with the coordinated international response.
Direct translation: this not is that a part of the contribution of the EU and this contribution must be in consistency with the reaction international coordinate
Our machine translation: This not is that a part of the contribution of the EU and this contribution must be in consistency with the international reaction coordinate.

2. French: Par contre, je ne peux pas accepter l'amendement 3.
Human translation: In contrast, I cannot accept Amendment 3.
Google Translate: For cons, I can not accept amendment 3.
Direct translation: by against I not can not accept the amendment 3
Our machine translation: To against I can not accept the amendment 3.

3. French: Ces auditions ont été une réussite retentissante pour le Parlement européen et, partant, pour les citoyens de l'Union européenne.
Human translation: These hearings were a resounding success for the European Parliament and hence for the people of the European Union.
Google Translate: These hearings were a resounding success for the European Parliament, and hence for the citizens of the European Union.
Direct translation: this hearing have summer a success resounding for the parliament European and part for the citizen of the Union European
Our machine translation: this hearings have summer a success resounding for the European parliament and thus for the peoples of the European union.

4. French: Honorable Parlementaire, je n'ai aucune peine à vous répondre, car je me suis moi-même permis de citer également l'importance de ce problème par rapport à l'élargissement.
Human translation: I have difficulty answering you, Mr Posselt, because I have already mentioned the importance of this issue in relation to the enlargement myself.
Google Translate: Honourable Member, I have no difficulty to answer, because I allowed myself to also mention the importance of this issue in relation to enlargement.
Direct translation: honorable parliamentary I not have no penalty to you answer car I me am myself allowed of quote also the importance of this problem by report to the enlargement
Our machine translation: Honorable parliamentary I have no penalty to you answer car I me am myself allowed of quote also the importance of this problem by report to the enlargement.

5. French: Au lieu de quoi, bien sûr, vous vous mêlez des affaires du Caucase du Sud, de la mer Noire, des affaires de régimes qui ne sont pas particulièrement stables et qui ne souhaitent peut-être pas cette ingérence.
Human translation: But instead, of course, you go dabbling in the south Caucasus, around the Black Sea, with regimes that are not terribly stable; with regimes that may not want us there.
Google Translate: Instead, of course, you mix business in the South Caucasus, the Black Sea, the business plans that are not particularly stable and can not be this interference wish.
Direct translation: the place of what well sure you you mix of business of Caucasus of south of the sea black of business of scheme which not be not particularly stable and which not wish perhaps not this interference
Our machine translation: Of place of what well sure you you mix of businesss of Caucasus of south of the sea black of businesss of scheme which be not particularly stable and which wish perhaps not this interference.

IV. Translation System

Pre- and post- processing strategies.
1. Language modeling
Our primary strategy was to implement language modeling to select the correct translation of every word. We used a trigram nltk language model, trained on the Brown corpus, with Laplace estimator and stupid backoff. We also took into account the order of translations in the dictionary (as the dictionary is sorted first by highest quality translation) by dividing all probabilities by the log of their index in the dictionary entry. This language model is used to evaluate the translation quality for every word and it is responsible for choosing the correct translation of each word. Therefore, it affects every sentence in the dev and test sets because it decides among all of the candidate translations of each word to produce the translation that we see.
example: choose “In Fact” instead of “In Reality”
dev set sentences: 1,2,3,4,5,6,7,8,9,10
test set sentences: 1,2,3,4,5

2. Adjective ordering
Our second strategy was to switch the ordering of adjectives in the French sentences since adjectives in French usually appear after the noun. However, in English, adjectives appear before nouns, so we switched the order of these word types in the translation.
example: grandma German → German grandma
dev set sentences: 2,3,4,6
test set sentences: 1 ("reaction international" → "international reaction"), 3 ("parliament European" → "European parliament" and "Union European" → European Union)

3. Removing extraneous articles
In French, an equivalent of the word "le," which means "the," is used after titles, but before a name. In English, we don't include this extraneous "the," so we removed "the"s when appearing after determiners or titles.
example: Mr. the President → Mr. President, a the crossroads → a crossroads
dev set sentences: 1,5,6,9
test set sentences: *none*

4. Correct prepositions
We also decided to correct the prepositions that differed from English, such as changing "of" to "to" before verbs and changing "NOUN of ADJECTIVE" to "ADJECTIVE NOUN" as we would see in fluent English.
example: time of more → more time
dev set sentences: 5,6,8
test set sentences: *none*

5. "Parce que"
Another strategy we implemented was to correct the phrase "parce que", which translates literally to "because that", to be simply "because." We believe that this is a consistent rule in the French language because it will appear every time the English equivalent "because" is used. In fact, a search of our corpus revealed that the phrase "parce que" appeared around 21,000 times.
example: Because that the grandmother → because the grandmother
dev set sentences: 5; consistent rule in language
test set sentences: *none*

6. Negations
We also corrected negations from French to English. French negations are two-parted, such as "ne peux pas," which in English would translate word-for-word to "not can not" since both "ne" and "pas" translate to "not". This is a consistent rule in the French language; the proper grammar is "ne VERB NEGATIVE-WORD", but in English, negations are always one-parted ("not VERB" or "VERB not"). As a result, part of our pre-processing entails remove one of the "not"s. While this only appeared once in the dev set, it appeared three times in the test set. We figured that this was just a quirk in our dev set, which is why we chose to implement it. EVERY negative sentence in the French language will be affected by this and negatives are a common structure in French, just as they are in English. A direct translation would get this wrong, but our system will be correct. In order to demonstrate this, I ran a grep for sentences with "ne" or "n'" and "pas" or "jamais" both in them (this will give a conservative estimate, because there are more words besides "pas" that serve the same function). 15% of all sentences in the corpus met these criteria, thus more than 15% of sentences in the corpus would have been affected by this rule.
example: Je ne peux pas → I not can not → I can not
dev set sentences: 6; consistent rule in language
test set sentences: 2 ("I not can not" → "I can not"), 4 (" I not have no penalty" → "I have no penalty"), 5 ("which not be not" → "which be not" and "not wish perhaps not" → "perhaps wish not")

7. Contractions
The presence of contractions in French causes what should be a two word English translation to be

incorrectly tokenized as one. We therefore expanded contractions out to their full form to allow for correct tokenization and translation.
example: l'échange → le échange → the exchange
dev set sentences: 4,5,6,8,9
test set sentences: 1,2,4

8. Plurals
Since we stemmed the words in the French sentences, this removes the pluralization from any nouns. To correct for this, we examined French phrases by looking for an enumerated list of words like “des”, “les”, “mes”, etc. to determine whether they were plural. If we found plural words, we then pluralized the English translation.
example: les Européens → <PLURAL>the</PLURAL> European → the Europeans
dev set sentences: 2,4,6,7,9
test set sentences: 3 (“hearing” → “hearings”), 5 (“business” → “businesss”)[2]

V. Comparison to Google Translate

1. Human translation: Mr President, thank you very much for your understanding.
Google Translate: Mr. Speaker, thank you very much for your understanding.
Our machine translation: Mr. presidents thank you many of your comprehension.

The Google translation of this sentence chose the word “Speaker” for Président, while our translation went with “president” and changed it to be plural. Since the human translation from Europarl also uses the word “President,” it would seem that our system chose a more accurate subject for the sentence but introduced additional inaccuracies with the incorrect pluralization. Other than this word choice, the Google and human translations are the same word-for-word, while ours fails to change the direct translation of “thank you many” to “thank you very much” and uses “comprehension” instead of “understanding.” This would suggest that the Google translation performed better than ours in this test case.

2. Human translation: In contrast, I cannot accept Amendment 3.
Google Translate: For cons, I can not accept amendment 3.
Our machine translation: To against I can not accept the amendment 3.

Neither the Google translation nor our translation of this sentence are very strong. First, Google translated “Par contre” to “For cons,” which does not make any sense as an English phrase. Our translation uses an equally poor translation of “To against.” The human translation makes it apparent that “Par contre” is a common phrase in French that is used to mean “In contrast,” but this is completely different from the direct translation of each of the words. Other than an extraneous “the” we failed to omit, our translation is the same as Google’s. Therefore, Google’s translation is probably slightly better, but neither is particularly good.

3. Human translation: These hearings were a resounding success for the European Parliament and hence for the people of the European Union.
Google Translate: These hearings were a resounding success for the European Parliament, and hence for the citizens of the European Union.
Our machine translation: this hearings have summer a success resounding for the European parliament and thus for the peoples of the European union.

The Google translation of this sentence appears to be very similar to the human one. Our translation, however, has several issues. First, the determiner “this” does not agree with the plural word “hearings”. Our capitalization scheme to capitalize the first letter of the sentence also failed and we didn’t swap the

[2] Note: This did properly tag it as plural. The reason that it says “businesss” instead of “businesses” is because our pluralizing algorithm was not quite correct

adjective “resounding” with the noun “success” as we did in other sentences due to an incorrect POS tag (it labeled “resounding” as a verb instead of an adjective). Furthermore, our translation has the odd word “summer” in it, again because it directly translated the word “été,” instead of using it as “had.” Other than this and the mistaken pluralization of the word “peoples,” our translation matches up with Google’s, but because of these errors, Google’s translation is better.

4. Human translation: I have difficulty answering you, Mr Posselt, because I have already mentioned the importance of this issue in relation to the enlargement myself.
Google Translate: Honourable Member, I have no difficulty to answer, because I allowed myself to also mention the importance of this issue in relation to enlargement.
Our machine translation: Honorable parliamentary I have no penalty to you answer car I me am myself allowed of quote also the importance of this problem by report to the enlargement.

The human translation of this sentence is different from the others in that it contains additional information that was not present in the original French sentence. For instance, the French contained the words “Honorable Parlementaire,” which would really translate to “Honorable Parliamentary,” as our system did. However, the Europarl translation inserted an actual name (“Posselt”) in place of “Parliamentary,” even though the name can’t be found anywhere in the sentence. The human translation is therefore not the best metric for success in this case.

The first way our sentence differs from Google is that Google chose to use the word “Member” rather than “parliamentary,” which does makes more sense and provides more fluency. Second, one of our most significant errors in this sentence is in the translation of “car je me suis.” This is another common French phrase that, when translated directly, means “car I me am,” but when used together means “because I.” Google was able to correctly translate this phrase, but our model was not. Our system also chose the wrong meaning of “peine,” selecting “penalty” instead of “difficulty” as the other translations did, simply because “difficulty” was not present in our list of possible translations for the word. Finally, our translation omits some important commas in the sentence, and translates “par rapport” literally as “by report” rather than the actual translation of these two words together, which is “relative.” Google’s translation is therefore more accurate than ours once again.

5. Human Translation: But instead, of course, you go dabbling in the south Caucasus, around the Black Sea, with regimes that are not terribly stable; with regimes that may not want us there.
Google Translate: Instead, of course, you mix business in the South Caucasus, the Black Sea, the business plans that are not particularly stable and can not be this interference wish.
Our machine translation: Of place of what well sure you you mix of businesss of Caucasus of south of the sea black of businesss of scheme which be not particularly stable and which wish perhaps not this interference.

This translation presents several interesting differences. The first is that our system translates “Au lieu de quoi” directly and literally, producing “Of place what.” It is traditionally used in French to mean something like “Instead,” which Google Translate gets correct. Our translation also does not properly interpret “vous vous” as a single “you.” It instead interprets it as two distinct “you”s as a result of translating each word literally. On the other hand, Google properly sees it as a single “you.” We also use the wrong preposition: using “of” instead of “in the,” and repeatedly interpreting “a” as “of” instead of “in.” Additionally, it doesn’t properly recognize “South Caucasus” and “Black Sea” as proper nouns, and as such doesn’t use our strategies to get the proper word order for the two, nor does it capitalize them properly. It also interprets “black” as a noun instead of an adjective, so it doesn’t switch “sea black” into “black sea.” Furthermore, the translation includes “businesss of scheme” instead of “business plans”; the former is translated almost directly while the latter uses more advanced features. However, neither of these is accurate; both systems interpret “régimes” as “scheme” or “plan” instead of “regime.” Our mistake here (and probably Google’s) is due to our language model. Additionally, neither system gets the conjunction correct; Google produces “and” and our system produces “which.” Both of the translations properly recognize the phrase “not particularly stable,” but neither of the translations do well at the end of the sentence. Google says “can not

be this interference wish," which isn't fluent English. We say "wish perhaps not this interference", which is slightly better, but still very awkward wording. Overall, Google's translation is better, although neither of them is very understandable.

VI. Error Analysis

1. Mistranslation of common phrases

Errors:

Sentence	Phrase	Our Translation	Human Translation
Test sentence 2	"Par contre"	"To against"	"In contrast"
Test sentence 5	"Au lieu de quoi"	"Of place of what"	"But instead"
Test sentence 5	"bien sûr"	"well sure"	"of course"

Although French and English share many common language structures and phrases, there are still numerous important French phrases that do not directly translate word-for-word to English. These phrases are most often contain prepositions or conjunctions. One systematic flaw in our translation system is in failing to preserve the meaning of French idioms that do not correspond word-by-word to English ones. For example, the French phrase "Par ailleurs" means "What is more" or "in addition", but we produced "To somewhere else". "Au lieu de quoi, bien sûr" means "instead, of course", but we produced "in place of what well sure".

One simple approach to improve this problem would be to amass a list of common French idioms and their English translations and to translate the idioms completely before attempting direct translation of the rest of the sentence. For example, we might have entries for "Au lieu de quoi" and "bien sûr" mapping to their respective translations. In preprocessing, we might check our phrasebook for these common phrases and group them together, so we translate them as one unit instead of distinct words.

A more complicated approach would be to incorporate each idiom into the language model and assess the relative likelihood of the words forming the English translation of the idiom vs. the words directly translated to French. For instance, with the example given here, the product of the probabilities of "in","addition", would be compared to the product of the probabilities "to","somewhere","else" and "in addition" would be more likely to be selected.

2. Subject Verb Agreement

Sentence	Phrase	Our Translation	Human Translation
Test sentence 5	"ne sont pas"	"which be not"	"which are not"
Dev sentence 6	"qui établit"	"which establish"	"which establishes"

One of the most obvious flaws in our translation is the poor subject-verb agreement. For example, our machine translation for test sentence 5 generates the phrase "which be not". This is not fluent English; the correct phrase would be "which *is* not". In the dev set, our translation generates "they…eats", while it should generate "they...*eat*." Our system design does not robustly check the subject of the clause and make the verb match it. Instead, it chooses the most common one.

There are two potential ways to fix this. One way would be a stronger language model. For example, a good language model trained on a robust corpus would probably be able to detect that “which be” is a much less common bigram than “which is”, and thus choose the word “is” over the word “be”. However, this would not handle all cases. For example, in our dev set sentence “Parce que la grand-mère allemande est diabétique, **tous les Européens** doivent-ils **cesser de manger** du sucre?”, the subject is “tous les Européens”, which translates to “all Europeans”, and the verb phrase is “cesser de manger” or “stop eating”. We translate “it manger” to “eats” instead of “eating”. A simple N-Gram model won’t account for this well, because the subject and verb are both many words long, and they are not right next to each other. To account for this with a language model, we would need something that would be able to look farther ahead.

A second and probably more feasible way would be to do explicit checking of the type of subject and verb. For example, the infinitive “to be” can be conjugated to “is”, “am”, and “be”, and “to have” can be conjugated to “has” or “have”. Just checking the tense wouldn’t be enough, because all of these translations are in the same tense. Instead, one step might be to check the “plurality” of the subject. For example, if we see the word “they” followed by a present-tense form of “to have”, we could determine that we should use “have” rather than “has”.