-Independent Work Report Spring 2014-

Sentiment Analysis of Foreign Documents to Detect Translation Bias: Misconstrued Communications in the European Parliament

John Whelchel & Christiane Fellbaum
Princeton University

Abstract

Modern translation studies neglect to try and quantify the effects of individual translator and interpreter biases, deliberate or inadvertent, in their work. Translators working in the diplomatic sphere are expected to produce translations of the highest quality so that no accuracy is lost. This paper shows that there is a demonstrable, quantifiable change in mood and sentiment across translations in documents from the European Parliament. Consequently, there is a lack of accuracy and loss of meaning between political groups and nations, which may cause problems or conflicts in the future due to the critical nature of diplomatic translation.

I. BACKGROUND:

Historically, the practitioners of translation and interpretation changed the content of a message they were translating if it served some ulterior motive. Dragomans in the 17th and 18th century Middle East would translate the haughty language of their leaders' communications to allied powers into humbler messages with more equal terms. Conversely, the messages they received would be made more supplicant and subservient in nature when rendered back in return to their masters. This was to both protect themselves ("don't shoot the messenger!") and to maintain good diplomatic relations between the varying nations and cultures. This rather blatant editorializing was not ideal in such a crucial job, and soon the world powers had their own translators beyond the dragomans (Lewis 2004).

Modern day simultaneous interpreters are often faced with similar problems due to the time sensitive and sometimes dangerous nature of their work, and their craft has been critically studied thoroughly in the academic community. "One of the key skills of the simultaneous interpreter is decisiveness: there is simply no time to weigh the merits of variant translations." states Edwards (2001). In the peacekeeping world there is a strong tension between accuracy and safety for all involved: "Errors in translation of peacekeeping negotiations can have a dramatic and costly impact on international missions." Ting Guo (2014) states, "When facing extreme situations such as wartime, interpreters can actively use their accumulated capital to negotiate benefits beyond the interpreting situation and protect themselves." Lynn Visson (2013), a UN interpreter, acknowledges this distinction as well: "The translator has time to change, edit, and refine his text...[an interpreter] only has the words and phrases in his brain to rely on [under pressure]."

Specifically, Visson also acknowledges that because of the intense pressure while interpreting, a diplomatic "bureacratese" has developed in the UN to simplify the decisions interpreters make by giving universally safe alternatives. "*Happy*" becomes "*satisfied*", "*theft*" or "*embezzlement*" becomes "*failure to ensure compliance with proper accounting and auditing procedures in the handling of financial resources*." This almost comedic couching of language is of debatable necessity in the sensitive, time-pressured world of the UN and other diplomatic interpreters.

However, these problems and choices have not only occurred during interpreting, but also through translations, despite the surplus resources and time available to translators. The Balfour Declaration is an infamous example of a mis-translation gone wrong with severe consequences in the diplomatic community, despite the lack of time pressure on translators. Chesterman (2014) discusses Robinsons *Translation and the Problem of Sway*, "sway" meaning "ways in which translation decisions and interpretations are influenced, consciously or unconsciously, by collective pressures". This "sway" can be detrimental to accurate communication, and translators without the same pressures as interpreters should theoretically be less liable to have sway present in their work. There is also a danger of sway being magnified as translations continue. There is a common practice of using one or more common languages as “pivot” languages, middlemen of sorts, when translating between two less common languages. Given the prevalence of this practice, bias could be magnified and distort true communications between parties and nations even more. As Chesterman and Báhegyi (2012) also explain, there is a strong distinction between intentional and inadvertent bias. Naturally, inadvertent bias as a subconscious process can not be easily identified, but its effects can be seen, thus it is vitally imperative to accurately translate diplomatic documents, as mistakes or subtle shifts of word choice and tone can have large effects that are amplified later.

As demonstrated, translation studies have been thorough in regards to the possibility and causes of bias in interpreting, yet more cursory in regards to translation. In addition, most of the evidence is first-hand and anecdotal, instead of quantitative. This paper uncovers and quantifies a similar bias that is distorting the translated works of diplomatic relations between nations. Despite the fact that translators have much more time and many more resources, they are either subconsciously or consciously subtly editing the documents they receive in order to make them more palatable to the diplomatic world, losing accuracy and exact meaning in the process. Using documents from the European Parliament, I found that on average the translations across the European languages under study in this paper (English, Spanish, Italian, German, and French) skewed towards neutral, and the effect was slight but significant. Typically, government documents are neutral overall, but by using a large parallel corpus, small linguistic choices that temper positive or negative language made apparent subtle biases on the large scale.

II. METHODOLOGY:

SENTIMENT ANALYSIS

Sentiment analysis is a technique used to obtain a sense of a document's or sentence's overall mood or character. There are many basic varieties of sentiment analysis, each with strengths and weaknesses (Kaur, Gupta 2013). The most simple and obvious is a “subjective lexicon”, also known as the “bag of words” approach, in which individual words are manually assigned positivity/negativity (or other) ratings by humans and then applied to sentences and documents as a whole. While superficially reasonable, this method is not always very effective. Consider the sentence, “I am not ecstatic”; the inversion effect of words such as “not” and other qualifiers are completely ignored in this approach. Another approach tries to mitigate this somewhat by going beyond one word at a time by using n-gram modeling. Other approaches use large amounts of machine learning via annotated training and test data sets. For this analysis, a machine-learning algorithm pioneered by Stanford researcher Richard Socher (Socher et al. 2013) was used. The main advantage of the algorithm is its ability to mitigate the disadvantage of the “bag of words” model, namely the inability to account for context.

This algorithm uses a “recursive neural tensor network” that builds a rating for each sentence by starting from the constituent words, creating a tree structure made of sub-phrases that combine the lower level leaves of words into larger and larger groups, accounting for changes in context as the tree is built until an overall rating is obtained at the root (Illus. 1). Consequentially, training data for the model requires data of various sizes, from individual words to entire sentences.

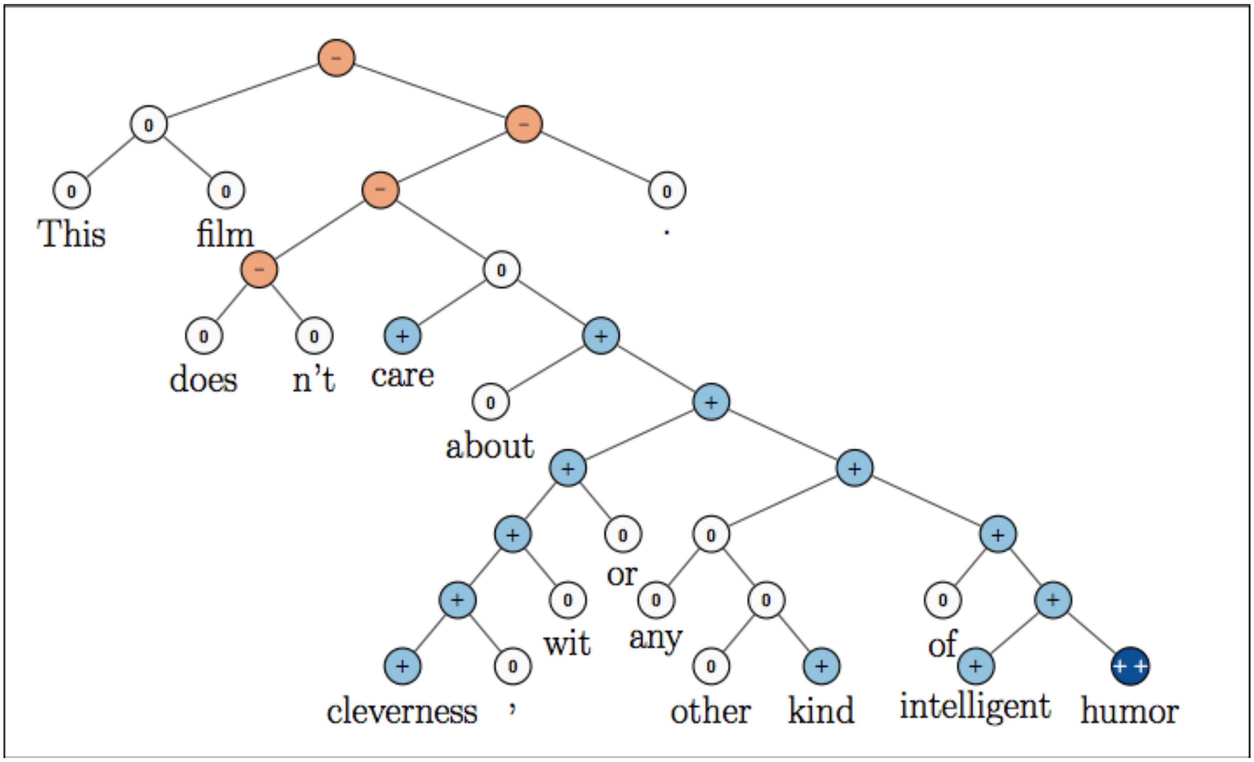

Illustration 1: Example of algorithm's tree-based approach that takes context into account. The root node is the overall rating for the sentence (Socher 2013).

DATA COLLECTION:

Thirty documents were collected from the European Parliament's online database. These documents came from three genres in the database: "Social Questions", "Politics", and "International Relations," and are from within the past eight years. Each document was downloaded in five languages, English, German, Spanish, French, and Italian, to create a large parallel corpus. They were converted to plain text and stripped of meta-tags and formatting in order to leave only phrases and sentences. There were six originals in each language among the thirty documents, each also having four translations. There were roughly 400,000 words for each language in total.

legislation, proposing that manufacturers should be required by technical means to reduce emissions to 130g CO2/km by 2012, with various complementary measures being introduced to ensure that overall emissions did not exceed 120g CO2/km by this date.

The 2012 target date is something of a chimera. It reflects various political decisions, most recently adopted unanimously by the Council in June last year, that have paid little heed to the actual situation. The Commission has tried to bridge the gap between desire and reality by proposing that the 120g CO2/km target should be achieved with the assistance of a range of 'complementary measures,' such as the use of biofuels, that could be taken into account in addition to technological improvements, even though these are likely to be affected by such variables as driver behaviour and the availability of alternative fuels.

The rapporteur believes that car makers should be given absolute certainty about the technical standards they are required to meet, and while the potential exists for introducing and incorporating some values for the CO2 reductions that can be achieved through use of

Text 1: Example excerpt from an English European Parliament Document

To collect data to use for the sentiment analysis, I solicited Princeton University students' assistance. I found three native speakers in each of the five languages (except for two from Germany), with a preference for those actually from Europe when possible (as opposed to South American Spanish speakers, for example). Each individual was given a small computer program that allowed them to rate a list of phrases and sentences on a scale from 1 to 5, with 5 being the most positive, 1 being the most negative, and 3 as neutral (Illus. 2). At any point in time, they could save their progress and quit to continue later, and they could also step backward to modify ratings as they saw fit (Illus. 3).

Many of the native speakers who helped me judge phrases observed that they rated a high proportion as neutral; this was not surprising, as diplomatic documents tend to contain regimented, formal language. The ratings given by each of three native speakers were averaged to get a score for each phrase or sentence, rounding *away* from 3 (neutral) when necessary in order to mediate for the large number of neutral scores. Example rated phrases from English include: “the International Criminal”, “the efforts made by China” and “bloodshed in Syria”.

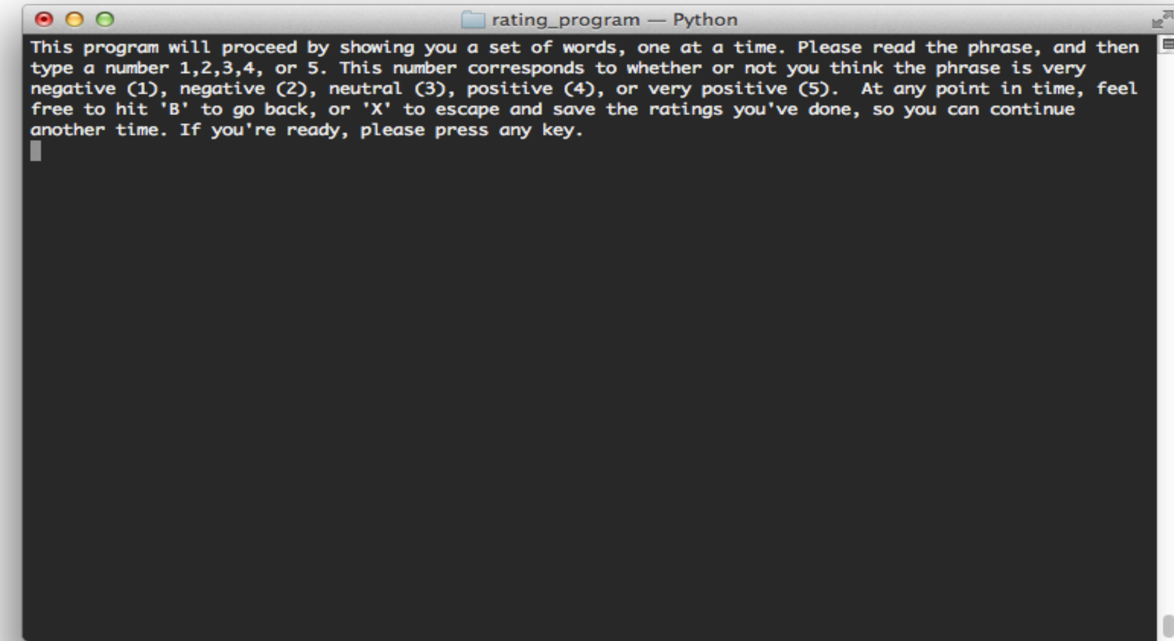

Illustration 2: Instruction screen for native speakers.

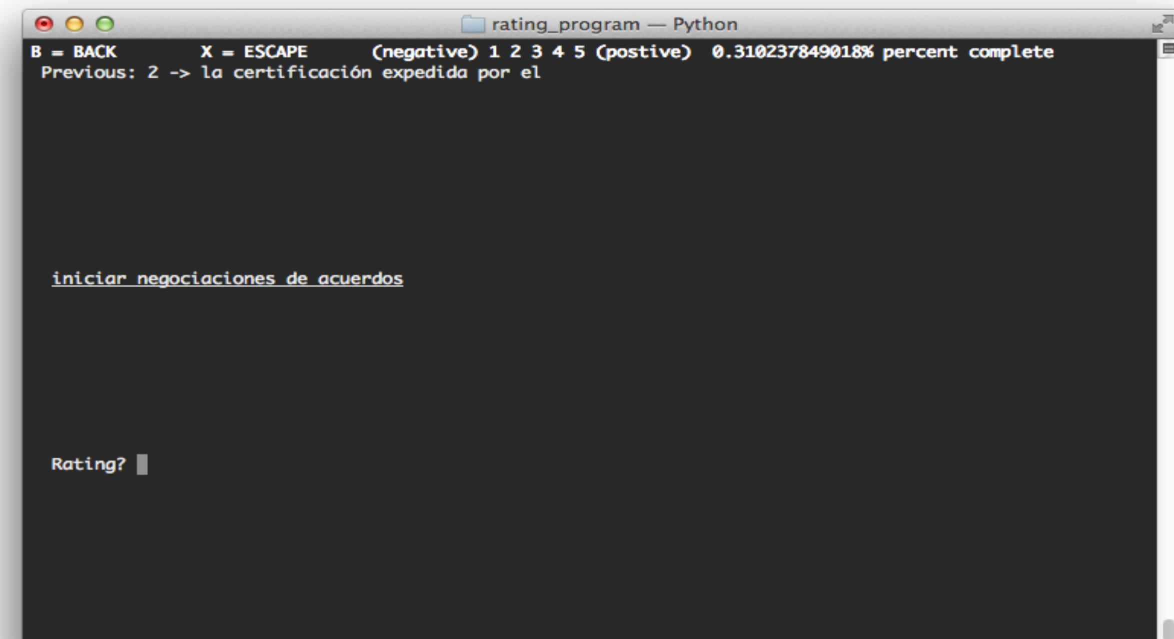

Illustration 3: Main rating screen with Spanish examples

There were a limited number of phrases and sentences that could feasibly be demanded of each native speaker due to the time demands of judging vast numbers of phrases. One thousand ratings per person was set as the target number. To create the datasets, all of the documents were converted to .txt files and stripped of miscellaneous meta-information and obvious non-sentences, such as solitary dates, document numbers, etc. They were then parsed into sentences, and collated into 5 large files, "sentences.txt", one file per language. Each was shuffled, and then the first 100 sentences were separated. From these 100, I created all possible 3-grams, 4-grams, and 5-grams. These were then shuffled, and the first 800 were removed. These 800 became the base of the judging for each language. To this 800 was added the 100 original sentences, in addition to the 100 most frequent words in the corpus for that language, less stop words.[1] Lastly, I removed any duplicates. Each final set of phrases and sentences to be weighted was roughly 970 items. These were each then shuffled (so as to avoid possible proximity bias between identical sequences of words and phrases given to the students) and given to the three native speakers in each language to rate. Given Zipf's Law[2] of natural language frequency, which in the case of some English corpuses might mean 135 words account for 50% of the word usage, I estimated that the top 100 words (less stop words) would account for ~30% of unique words (Fagan 2010), and thus were worthy of including in the data sets in order to give ample ratings for the leaves of the trees in the machine learning algorithm. Phrases and words without ratings received the rating of the overall sentence.

The sentiment analysis model was trained on the ratings given by native speakers, and then run on all documents in each language. To train, the ratings given by the native speakers were averaged, and then broken up into a "dev-set" and a training set. The "dev-set" allows the

---

[1] Stop word lists came from the following sources:
Eng.: http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
Ger.: http://snowball.tartarus.org/algorithms/german/stop.txt
Fr., Sp., and It.: http://members.unine.ch/jacques.savoy/clef/${language}ST.txt

[2] Zipf's Law states: the frequency of a word is proportional to the inverse of its frequency rank.

algorithm to compute hyper-parameters during training (Socher 2013). The "dev-set" was roughly 200 phrases, while the rest was reserved for training. This proportion was comparable to that used in Socher's work.

For each document, a score was obtained for the original-language version by judging every sentence, and then averaging them all. This was then compared against the four translations, each rated in the same manner, to get an overall perspective on possible differences or biases.

III. RESULTS:

In total, the student native speakers made 13,681 ratings. The ratings distribution is as shown in Fig. 1, highlighting the general bell curve shape of each bar cluster in the graph, shifted slightly towards positive. The vast majority of weightings fall in the first middle three categories. This is due in part to the rounding away from neutral in an effort to increase the sensitivity of the model by avoiding a surplus of neutral ratings. Without the rounding, it is likely there would

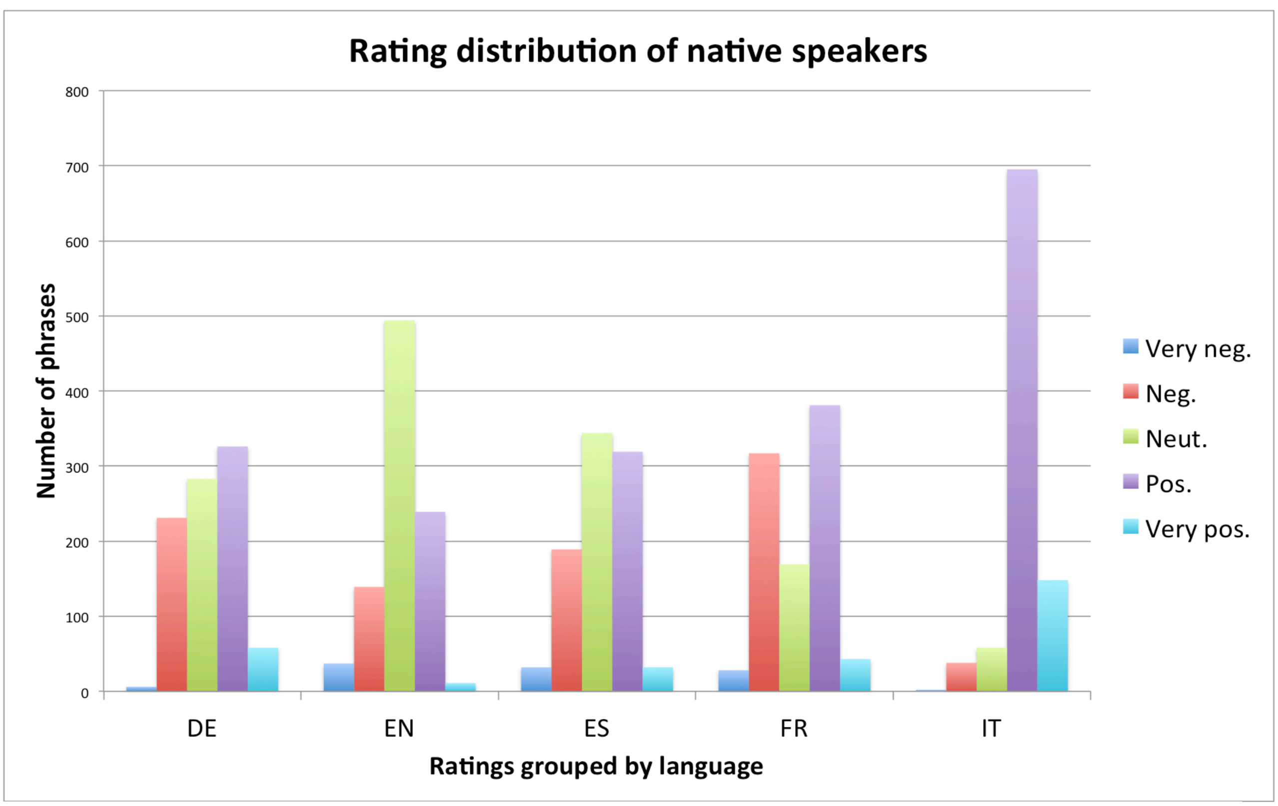

Figure 1

Category	DE		EN		ES		FR		IT	
	(trans)	(orig)	(trans)	(orig)	(trans)	(orig)	(trans)	(orig)	(trans)	(orig)
Very neg.	0.001	0.002	0.041	0.067	0.247	0.358	0.002	0.002	0.0001	0.000
Neg.	0.681	0.707	0.240	0.244	0.140	0.205	0.576	0.656	0.282	0.412
Neut.	0.063	0.066	0.387	0.228	0.190	0.134	0.011	0.012	0.028	0.040
Pos.	0.110	0.115	0.323	0.444	0.366	0.212	0.389	0.310	0.623	0.493
Very pos.	0.144	0.110	0.008	0.016	0.058	0.091	0.022	0.021	0.066	0.055

Figure 2

	DOC	Orig.	Tran.
EN	1	3.119	2.993
	2	3.207	2.988
	3	3.097	2.931
	4	2.987	2.948
	5	3.051	2.976
	6	3.164	2.868
FR	7	2.824	2.987
	8	2.837	3.040
	9	2.875	3.017
	10	2.883	3.056
	11	2.660	3.013
	12	2.739	3.000
DE	13	2.667	2.799
	14	3.003	3.018
	15	2.863	3.008
	16	2.432	3.100
	17	2.733	3.006
	18	2.674	3.003
IT	19	3.322	2.867
	20	3.244	2.887
	21	2.880	2.823
	22	3.659	2.781
	23	3.310	2.880
	24	3.721	2.836
SP	25	2.367	3.087
	26	2.535	2.971
	27	2.573	3.043
	28	2.608	3.072
	29	2.340	3.106
	30	2.543	3.018

Figure 3

have been almost no “very negative” or “very positive” phrasal ratings at all. Note the outlier in the Italian positive bar (and somewhat in French with neutral). Italy’s positive bar accounted for over 70% of the Italian judgments.

Using this data, five sentiment models were created for each language; each model was trained for forty rounds. These models were then used to evaluate each sentence in every document. The sentences were averaged to get a document score, and then computed as normalized percentages as well, allowing comparisons between the originals and the translations.

Fig. 2 details the expected percentage of ratings across the five languages, separated into translations and originals. The percentages for ratings that were not neutral were expected to decrease, while the opposite was expected in the neutral category. Fig. 2 highlights squares to indicate where this skewing towards neutral occurred. Fig. 3 showcases this comparison on a document basis using the document score for the thirty documents. In the case of translations, the document score of each translated version was averaged together. It is clear at a cursory glance that virtually all of the translated document scores are closer to three (neutral) than the originals.

Fig. 4 unpacks the data in more detail, breaking the thirty documents up by language. Each set of six documents compares the prospective ratings between originals and translations by sentence percentages. It is apparent that when grouped by language, the documents either tend to

Orig.		Orig.	Tran.	Orig.	Tran.	Orig.	Tran.	Orig.	Tran.	Orig.	Tran.	**Neutral motion**
	DOC	Very N.	Very N.	Neg.	Neg.	Neut.	Neut.	Pos.	Pos.	Very p.	Very p.	
EN	1	0.074	0.107	0.243	0.375	0.181	0.053	0.494	0.345	0.008	0.119	**-0.256**
	2	0.074	0.073	0.174	0.399	0.223	0.057	0.529	0.410	0.000	0.061	**-0.332**
	3	0.057	0.085	0.276	0.423	0.197	0.053	0.452	0.354	0.018	0.085	**-0.288**
	4	0.084	0.103	0.252	0.389	0.265	0.064	0.389	0.346	0.009	0.098	**-0.403**
	5	0.060	0.098	0.274	0.402	0.239	0.044	0.410	0.341	0.017	0.116	**-0.392**
	6	0.053	0.107	0.217	0.424	0.286	0.055	0.402	0.323	0.042	0.091	**-0.462**
FR	7	0.000	0.109	0.614	0.332	0.006	0.115	0.324	0.352	0.057	0.092	**0.218**
	8	0.005	0.093	0.587	0.326	0.005	0.121	0.372	0.366	0.031	0.093	**0.232**
	9	0.000	0.111	0.588	0.325	0.008	0.103	0.346	0.356	0.058	0.104	**0.190**
	10	0.015	0.117	0.555	0.285	0.007	0.120	0.380	0.384	0.044	0.095	**0.226**
	11	0.001	0.065	0.669	0.343	0.012	0.175	0.302	0.348	0.015	0.069	**0.325**
	12	0.000	0.125	0.638	0.321	0.021	0.099	0.303	0.339	0.037	0.116	**0.156**
DE	13	0.039	0.153	0.627	0.337	0.059	0.078	0.176	0.422	0.098	0.010	**0.039**
	14	0.000	0.100	0.557	0.313	0.077	0.113	0.173	0.416	0.193	0.057	**0.073**
	15	0.000	0.123	0.611	0.297	0.064	0.099	0.175	0.413	0.149	0.069	**0.069**
	16	0.000	0.058	0.768	0.282	0.079	0.190	0.106	0.442	0.047	0.028	**0.222**
	17	0.004	0.063	0.688	0.340	0.049	0.160	0.090	0.401	0.169	0.036	**0.221**
	18	0.000	0.106	0.681	0.333	0.052	0.076	0.178	0.422	0.089	0.063	**0.049**
IT	19	0.000	0.057	0.355	0.425	0.028	0.167	0.558	0.299	0.060	0.053	**0.278**
	20	0.000	0.061	0.380	0.416	0.039	0.157	0.538	0.308	0.043	0.058	**0.237**
	21	0.000	0.063	0.560	0.404	0.048	0.205	0.344	0.303	0.048	0.025	**0.314**
	22	0.000	0.113	0.201	0.425	0.037	0.107	0.665	0.276	0.098	0.079	**0.141**
	23	0.000	0.069	0.346	0.410	0.055	0.157	0.542	0.301	0.057	0.063	**0.205**
	24	0.000	0.105	0.190	0.411	0.031	0.122	0.646	0.267	0.133	0.095	**0.182**
SP	25	0.360	0.021	0.273	0.415	0.094	0.088	0.187	0.408	0.086	0.068	**-0.012**
	26	0.307	0.011	0.198	0.452	0.188	0.117	0.267	0.395	0.040	0.025	**-0.143**
	27	0.302	0.009	0.188	0.430	0.188	0.112	0.281	0.407	0.042	0.043	**-0.152**
	28	0.365	0.023	0.162	0.418	0.122	0.084	0.203	0.416	0.149	0.059	**-0.075**
	29	0.422	0.020	0.189	0.418	0.126	0.068	0.150	0.422	0.112	0.072	**-0.117**
	30	0.315	0.008	0.228	0.436	0.120	0.112	0.272	0.415	0.065	0.028	**-0.015**

Figure 4

move as a group either towards or away from neutral, as shown in the final column. The middle three languages, French, German, and Italian, seem to have the most motion towards neutral when they are the original language of the document. The average motion towards neutral per document is 2.43% across all languages.

IV. ANALYSIS:

There is a demonstrable and clear correlation of decreasing non-neutral ratings and increasing neutral on average when European Parliament documents are translated. Looking at the documents on average, Fig. 5 shows that virtually every document has an absolute value score farther from neutral than the corresponding translation. Additionally, the general spikes and curvature of both line graphs mirror each other.

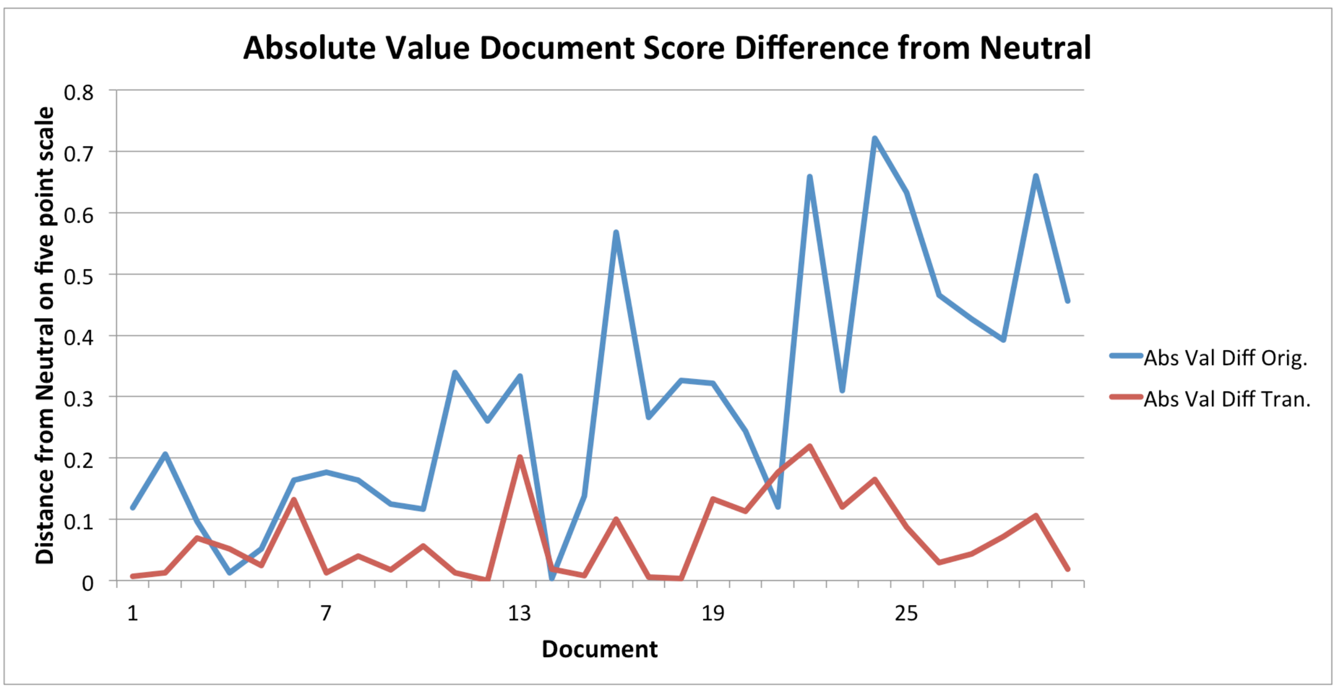

Figure 5

Documents 5, 13, 15, and 19-25 fall and rise together, while the rest correlate obliquely. This indicates there is a pattern of tempering language when translating, though it is virtually undetectable when the original document is very neutral on its own (Documents 3 and 14, for example). The original document farthest from neutral has a document score of ±.7, while its translations have an average sentiment distance of .15 from neutral, meaning that the original document is more than half a data point removed from its original mood, greatly altering the timbre and feel of the original document. The graph becomes more interesting when the absolute value is removed, as the peaks and troughs that mimic each other are in fact inverted (Fig 6). This seems to indicate that not only is bias occurring, but it is in fact over-correcting beyond neutral to compensate for the positive or negative language! The document farthest from its translation in Fig. 5 has now increased in distance from its translation almost one whole sentiment value, as if going from directly from positive to neutral or negative to neutral. The two lines always cross each other near the line of neutrality as shown by documents 6, 18, and 24. This shows quite strongly that documents on average do change in mood when translated, and they approach neutrality instead of positivity or negativity. The average document score for the

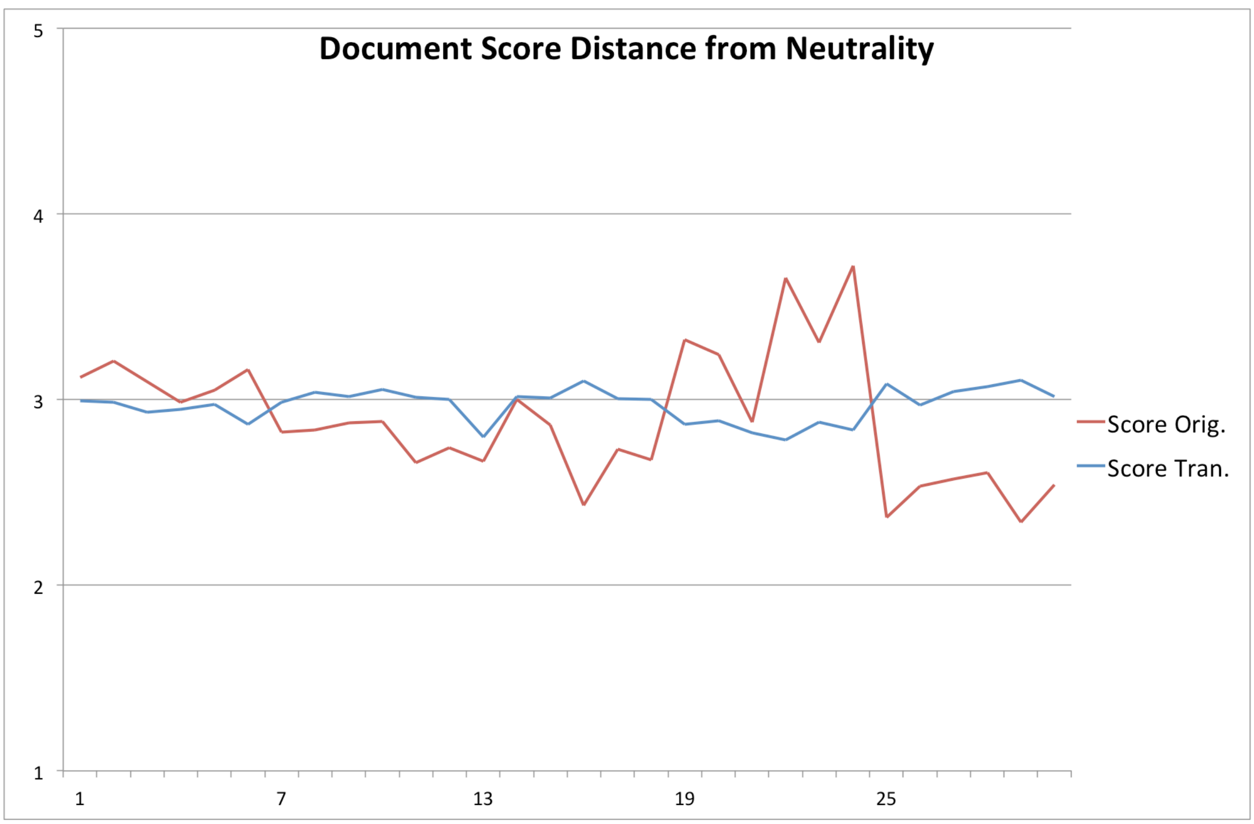

Figure 6

original documents was 2.900, while the average document score for the translations was 2.970, closer to the neutral 3 and thus corroborating the claim.

By looking at the data on a more individually linguistic level, a different pattern appears. Figure 7 breaks up the data by language to compare how much each moves towards neutrality when translated.

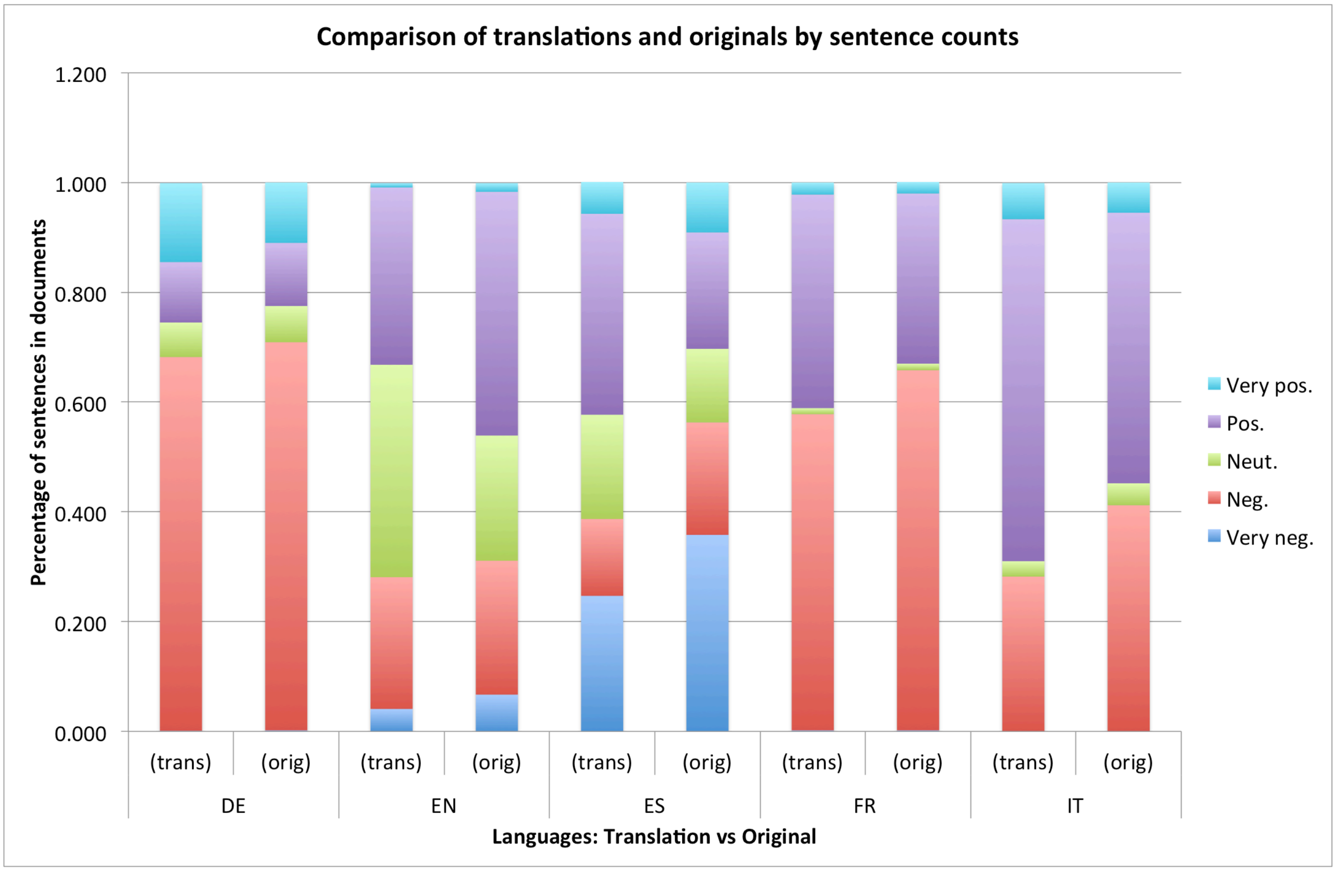

Figure 7

The immediate striking aspect of the column graph is that not every language actually results in a direct increase of neutral ratings (with a corresponding decrease in others, as the documents' scores are zero-sum). English seems to be tempered very effectively, as its green neutral bar increases the most when translated. Spanish also seems to do the same, while the rest appear rather inconclusive.

It is difficult to determine exactly why this occurs. Italian and French seem especially similar between translations and originals in terms of distribution of neutral sentences and non-neutral sentences within Fig. 7. It is possible that the rating data given for Italian was not of a high enough quality or of a large enough quantity, given the strange distribution in takes in Fig. 1. The general shape of the ratings for Italian does not match any of the other languages and is disproportionately high in the "positive" category. Despite Italian's movement away from neutrality, the data shows that there was a 16.4% total increase in neutrality, with a corresponding 8.5% decrease, resulting in a net 7.9% increase across all the languages.

Discounting Italian, there is some pattern in the other four, notwithstanding the fact that their proportions of neutral ratings do not all increase. In fact, each language but Italian moves centrally on average, as can be seen by the neutral bars moving towards 0.5 for the first four. After being translated, the ratio of positive and negative statements moves closer to one, resulting in a net increase in average neutrality (though not necessarily actual! A document can be very positive and very negative yet neutral overall). The neutral section of German centers marginally, while Spanish moves significantly more towards the middle, increasing neutrality, though the actual number of neutral sentences only increases by a small factor. This is a strong indicator that even on a linguistic level, there is a correlation between translation and movement towards neutrality. It is unclear why there are differences between the languages in terms of how the neutrality is expressed. It could be attributed to the methodology, the language itself (semantic patterns that differ and syntactic quirks mishandled by the algorithm originally intended for English), or even possibly patterns of specific translators. It is possible that schools of thought and training methodologies for translators across nations and cultures differ, resulting in different levels and manifestations of neutrality bias. As mentioned above, the average motion towards neutrality per document was 2.4%, while the average motion towards neutrality per language was 7.9%. This seems to indicate that language dominates subject matter, and that certain languages (notably English, here) experience more neutrality bias than others.

Lastly, Fig. 8 shows an interesting line graph that seems to defy the previous one. The x-axis marks documents in groups by language and indicates by what percentage they either moved towards or away from neutrality, with a positive value meaning toward. It shows that documents that originate in English and Spanish move the least towards neutrality (they in fact move away!), while documents in French, German, and Italian become more neutral. However, this actually validates the previous claims. If a document originates in English, it experiences the least "sway", since English translations move the most towards neutrality and it of course is not being translated into English. Conversely, documents originally written in French, German, or Italian will experience great movements toward neutrality because of the neutral bias caused by the English and Spanish translations (and the other two to some extent).

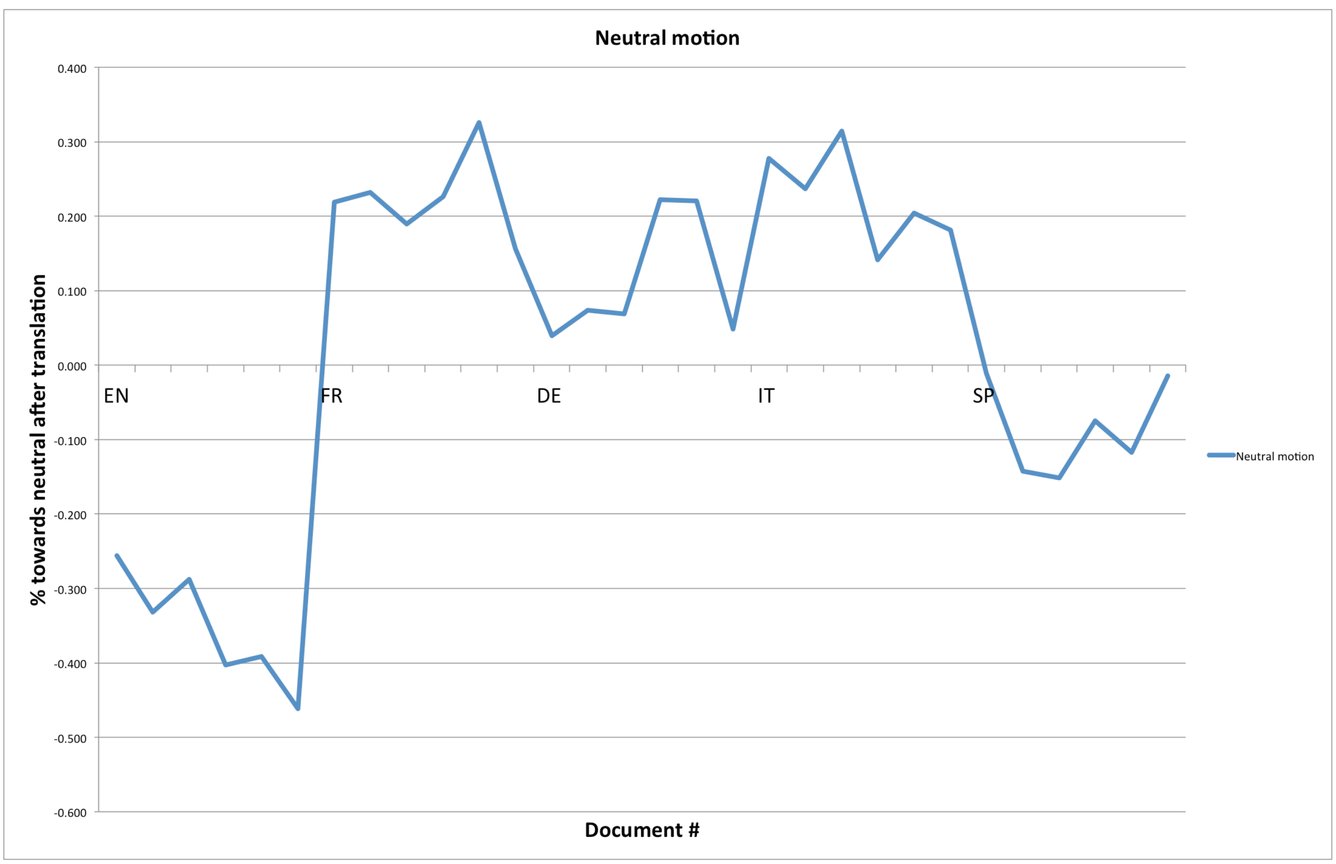

Figure 8

V. CONCLUSIONS:

The data clearly shows that neutrality bias is present and quantifiable in the work of diplomatic translators at the European Parliament, despite the less stringent time pressures and ample references and resources available. It is unclear whether or not this can be attributed to deliberate or inadvertent bias. As mentioned before, Visson discussed a UN interpreter "bureaucratese" which bowdlerizes and hides frank words in vague, politically correct statements. It is possible that across many international organizations that demand translators, this is also the norm. If so, there is a tension between translators who strive to soften the documents that they are translating, and the desires of perfect accuracy of those for whom they translate. By changing the sentiment of documents slightly (or significantly, as the data suggests sometimes occurs), everyone who interacts with that translation is left at a disadvantage by being given incomplete information. If translators are being explicitly or implicitly asked to temper the language of diplomatic documents, something should be done to maintain the fidelity of those documents as well.

Alternatively, it is possible that many translators strive to convey how something was said in addition to what was said, and are unaware of the subtle neutrality bias that is occurring. If so, this is exceptionally remarkable, as the data suggests translators are over compensating for negative and positive documents unwittingly (Fig. 6). In this case, this paper will shed light on the problem and hopefully encourage translators to approach their own work more critically, so that their translations portray sentiment very accurately. In this modern era where the UN and European Parliament act as negotiators and brokers between hundreds of nations, accuracy should be of highest priority.

VI. FUTURE WORK:

Given the confirming nature of the results, I would like to test the experiment again using a much larger training data set, predominantly by getting more native speakers and consequently more phrases rated. There was an abundance of documents and sentences to be weighted, unfortunately only a small subset could be weighted by only a handful of people in each language. Perhaps some online micro-payment system like Mechanical Turk would be advantageous for this, although it would be difficult to verify if people are truly native speakers, as this is important for maintaining cultural distinction and shared sentiments. My data did not seem to be very equal across languages or entirely consistent. A handful more native speakers along with more phrases would have increased the fidelity of the data analysis, although the correlation was still clear.

It would be important to also move beyond languages in the European family, especially the closely related ones studied here. The next step might be to study the UN and their documents, as they are available in vastly different languages (Arabic, Mandarin, etc.). As the UN covers a more global sphere, there are more cultural and societal differences between diplomatic nations that might be more incentive to engage in neutrality bias, resulting in more conclusive or even extreme “sway”. This study specifically focused on diplomatic documents that were likely to have ulterior motivations for neutral bias, notably potentially controversial or sensitive subjects such as “Social Questions” and “Politics”. It is not at all clear if this bias is as pronounced or even exists at all in the diplomatic documents of a more mundane nature. More research needs to be done to establish whether or not neutrality bias is truly pervasive, or is only systemic to sensitive subjects. If it were to be systemic to only controversial topics where it might behoove a translator to temper language, this would be some evidence for deliberate as opposed to inadvertent neutrality bias.

It would be fascinating to try and see if certain translators exhibit more translation bias than others. Fig. 7 indicates there is variance across languages, and this could be due to the styles of individual translators. Essentially, the training of translators within one country and between countries cannot be perfectly standardized. Some translators might not be discouraged (or perhaps they are even encouraged) to engage in neutrality bias as their role as a translator. Unfortunately, the data I collected had no translator data attached to it, although I reached out to the European Parliament for that information. Yet given the clear nature of the results, it might be possible to obtain it for future experiments, or from other data sets. With translator data, even more variables could be probed, such as how a translator's experience, background, and nationality affect the extent to which they exhibit neutrality bias.

Though neutrality bias may result in more cordial relations between powers, the possibility of miscommunication (as well as the fact that perhaps deliberate miscommunication is occurring at all!) should cause great pause in the international relations community. As a global society we are only able to function, cohabitate, and share the resources of our planet if we are able to communicate effectively. Translators in some of the most important realms of international public policy are modifying the nature of translated documents through calculated word choice in a way that changes the mood of communication. Time will only tell if this constant pipeline of modifications will result in another snafu such as the Balfour Declaration, whose ramifications still affect the Gaza Strip and Israel today. Thus more quantitative analysis needs to be done to determine the pervasiveness of this new neutrality bias so that our diplomatic relations will continue with accuracy and peace.

Bibliography

Bánhegyi, Mátyás. “Screening Political Bias and Reality in Media Translations.” *Translation Journal 16.* (accessed April, 2014)

Chesterman, Andrew (2014) Translation and the Problem of Sway, Translation Studies, 7:2, 224-227, DOI: 10.1080/14781700.2014.880367

Edwards, Victoria. "The Role of Communication in Peace and Relief Mission Negotiations." *Translation Journal* 6. (accessed April, 2014).

Fagan, Stephen; Gençay, Ramazan (2010), "An introduction to textual econometrics", in Ullah, Aman; Giles, David E. A., *Handbook of Empirical Economics and Finance*, CRC Press, pp. 133–153.

Kaur, Amandeep and Vishal Gupta. 2013. "A Survey on Sentiment Analysis and Opinion Mining Techniques." *Journal of Emerging Technologies in Web Intelligence* 5 (4): 367-371. http://search.proquest.com/docview/1500819461?accountid=13314.

Lewis, Bernard. *From Babel to dragomans: interpreting the Middle East*. New York: Oxford University Press, 2004.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts, “Recursive Deep Models for Semantic Constitutionality Over a Sentiment Treebank.” Presentation at the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, WA, October 18-21, 2013. http://nlp.stanford.edu/sentiment/.

Ting Guo (2014): Interpreting for the enemy: Chinese interpreters in the Second Sino-Japanese War (1931–1945), Translation Studies, DOI: 10.1080/14781700.2014.881302

Visson, Lynn. "Diary." *London Review of Books* 35 no. 21 (2013): 39, http://www.lrb.co.uk/v35/n21/lynn-visson/diary.

This paper represents my own work in accordance with University regulations:

X__John Whelchel__