Ana Daley (amd2yc@virginia.edu)
DS 5001
06 May 22

Lost in Translation:
An Exploration of Miguel de Cervantes Saavedra's Short Stories

Introduction

The general idea behind this project is the understanding that when words and more broadly, stories, are translated from one language to another, nuances can be forgotten and emotions overlooked. More simply, sentiments and meanings are "lost in translation". Knowing this begs the question, just how much is lost, and motivates the desire to analyze a text both in its language of origin and as a translation in order to compare the results. The specific goal of this project is to analyze the text from the short stories in Cervantes' Novelas Ejemplares, Exemplary Novels, in both their language of origin, Spanish, and in English. These results will then be used to compare the similarities and differences between the two versions.

Cervantes, a very highly regarded Spanish writer from the late 16th century, is the author of what many consider to be the first modern novel, Don Quijote. While this is what he is most well known for, he also authored 12 short stories that were published collectively between the first and second volumes of Don Quijote. These stories are the focus of this project as they have a wide range of topics and emotions, and as claimed by Cervantes, are the first short stories to be written in Spanish.

The Data

Novelas Ejemplares and its English translation were obtained from Project Gutenberg. The English translation was produced by Walter K. Kelly. Novelas Ejemplares was published in 1613 and contains the following 12 short stories: Coloquio que Paso Cipion y Berganza (Dialogue between Scipio and Berganza), El Amante Liberal (The Generous Lover), El Casamiento Engañoso (The Deceitful Marriage), El Celoso Estremeño (The Jealous Estramaduran), El Licenciado Vidriera (The Licentiate Vidriera), La Española Inglesa (The Spanish English Lady), La Fuerza de la Sangre (The Force of Blood), La Ilustre Fregona (The Illustrious Scullery-Maid), La Jitanilla (The Little Gypsy Girl), La Señora Cornelia (The Lady Cornelia), Las Dos Doncellas (The Two Damsels), and Rinconete y Cortadillo (Rinconete and Cortadillo). The data was tokenized and annotated using NLTK and Spacy libraries.

Analysis

After generating the data tables, the term frequency-inverse document frequency (TFIDF) scores were calculated and aggregated by each term. The associated bag of words (BOW) was done at the "book" or "story" level as these were short stories with no chapter distinctions. A subset of the nouns and adjectives in each corpus were sorted by TFIDF and the top 10 in each language were reviewed to see how well they overlapped.

The most significant nouns in English were: duke, grandfather, maestro, nag, tent, grandson, hospital, carrier, and brigantine and the most significant in Spanish were: duque, nave, dueña, bajá, alférez, esclaves, mantillas, bajel, and turcos. While the only noun that directly matches is duke (duque), there is also some overlap with ships/maritime words such as carrier and brigantine and bajel (vessel), nave (ship), and alférez (ensign). When looking at the adjectives, the most significant in English were: english, tunny, catalonian, crucifix, turkish, inviolable, primary, and unequivocal and the most significant in Spanish were: católico, jitana, inglesa, enojado, vieja, damascos, and polvos. Again, only one term directly overlaps, english (inglesa) but there are a few other associations with the terms, such as: crucifix with católico. Here we also see the term "turkish" as one of the most important English adjectives and that calls back to the term "turcos" in the significant Spanish nouns.

After looking directly at the nouns and adjectives deemed significant by their TFIDF score, dendrograms were generated as a means of evaluating how closely associated the stories were with each other and how this varied by language. These results can be found in Figures 1 and 2, below.

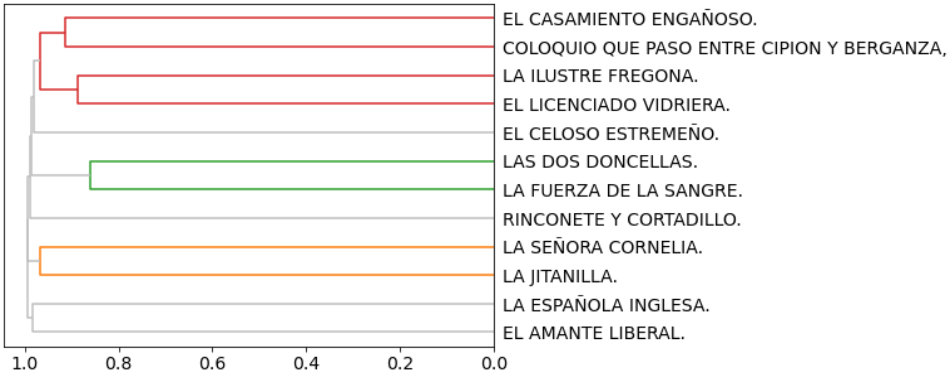

Figure 1. Dendrogram of Spanish Corpus

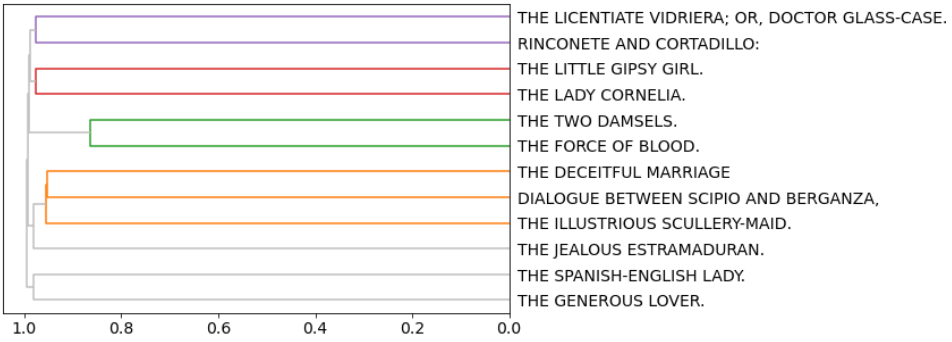

Figure 2. Dendrogram of English Corpus

In the dendrogram working from the Spanish terms, There is a direct grouping between The Deceitful Marriage (DM), Dialogue between Scipio and Berganza (SB), The Illustrious

Scullery-Maid (SM), and The Licentiate Vidriera (LV). A very similar grouping exists in the English Dendrogram with the exception of LV. In the English version, LV is grouped with Rinconette and Cortadillo (RC) whereas RC is more free standing in the Spanish version. In both the Spanish and English versions, the relationship between The Two Damsels (TD) and The Force of Blood (FB), The Little Gypsy Girl (LG) and The Lady Cornelia (LC), and The Spanish-English Lady (SL) and The Generous Lover (GL) are found. The Jealous Estramadauran (JE) has no immediate partner in either figure, but it is recognized as connected to the cluster of DM, SB, and SM in both. Thus, with the exception of RC, the clustering worked very similarly and overall, recognized the same relationships between the stories in both languages.

Principal Component Analysis was used with each language to see if there would be similarities in the identified components and the subsequent clustering of titles. The visualizations can be found below in Figure 3. The results of running PCA on the Spanish text are in the first row and the results from the English text are on the second row. The results can be seen full size in the accompanying Jupyter notebook titled, ANALYSIS.

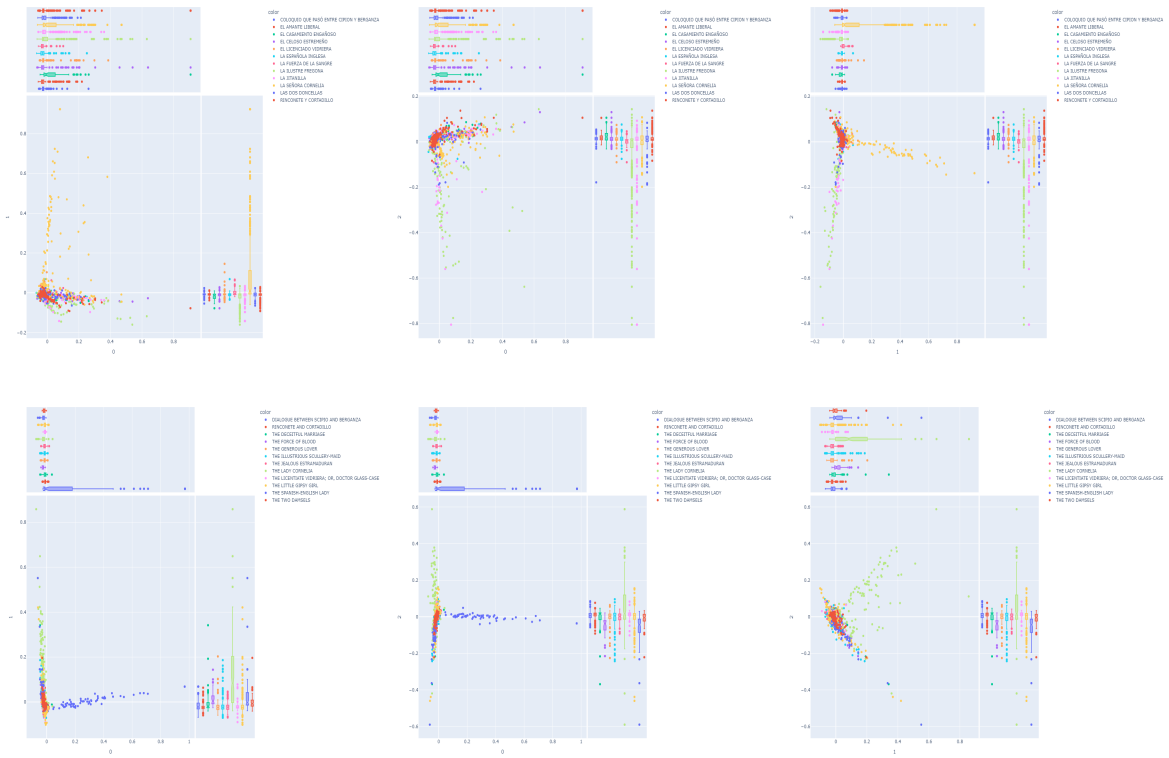

Figure 3. PCA by Paragraph and Language

While there are some similarities, such as seeing LC, SL and LG with the most spread while the rest of the short stories are more closely clustered, it is difficult to be able to easily glean comparisons between the two languages from these results. This is, in part, because the resulting principal components are most likely varied between the two, precluding any direct comparison.

As a main goal of this analysis is to understand the similarities and differences of the underlying tones and meaning of the texts in different languages, Latent Dirichlet Allocation was used to derive 5 topics from the texts and then evaluate which titles were most closely associated with each topic and determine which words were most relevant to each topic. See Figures 4 and 5 detailing which words were most relevant to each of the derived topics.

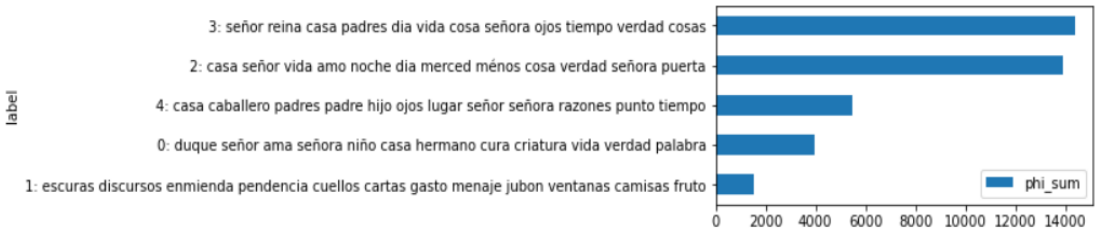

Figure 4. Most Important Words by Topic in Spanish

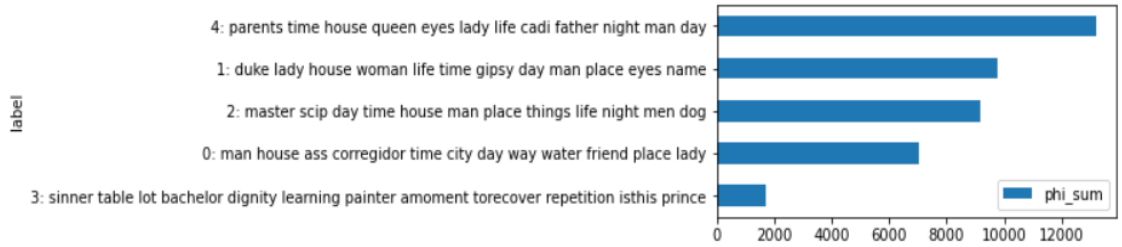

Figure 5. Most Important Words by Topic in English

The words associated with the most prevalent topic derived from the Spanish text are: man, queen, house, parents, day, life, thing, lady, eyes, time, truth, and things and of these 12 terms, 9 overlap with the most prevalent topic from the English text. Looking at the second most prevalent, the words associated with Spanish text are: house, man, life, love, night, day, mercy, less, thing, truth, lady, and door, of these 12 terms, only 4 overlap with the corresponding topic derived from the English text. Finally looking at the third most prevalent topic from each, the Spanish words are: house, knight, parents, father, son, eyes, place, man, lady, reasons, point, time; of these 12, 3 overlap. Therefore, with the exception of the main topic, the associated terms do not appear to be very similar and there seems to be more repetition of words between topics when reviewing the English terms.

Next, each title was associated with the topics as a means of understanding which topics are present in each story and in order to evaluate if there are similar clusterings in the stories and their topics. See the heatmaps in Figures 6 and 7.

	topic_id	0	1	2	3	4
title	chap_num					
COLOQUIO QUE PASO ENTRE CIPION Y BERGANZA,	1	0.000049	0.000049	0.999803	0.000050	0.000049
EL AMANTE LIBERAL.	1	0.000064	0.000063	0.000064	0.999746	0.000064
EL CASAMIENTO ENGAÑOSO.	1	0.000231	0.000228	0.000234	0.999075	0.000232
EL CELOSO ESTREMEÑO.	1	0.000084	0.000083	0.999664	0.000085	0.000085
EL LICENCIADO VIDRIERA.	1	0.000117	0.000114	0.000117	0.999536	0.000116
LA ESPAÑOLA INGLESA.	1	0.000071	0.000070	0.000071	0.999717	0.000071
LA FUERZA DE LA SANGRE.	1	0.000144	0.000142	0.000145	0.000145	0.999424
LA ILUSTRE FREGONA.	1	0.000057	0.000056	0.999773	0.000057	0.000057
LA JITANILLA.	1	0.000049	0.000048	0.000049	0.999806	0.000049
LA SEÑORA CORNELIA.	1	0.999667	0.000082	0.000084	0.000084	0.000084
LAS DOS DONCELLAS.	1	0.000081	0.000080	0.000081	0.000081	0.999677
RINCONETE Y CORTADILLO.	1	0.000088	0.000087	0.999650	0.000088	0.000088

Figure 6. Topic and Title Associations in Spanish

	topic_id	0	1	2	3	4
title	chap_num					
DIALOGUE BETWEEN SCIPIO AND BERGANZA,	1	0.000048	0.000048	0.999807	0.000048	0.000048
RINCONETE AND CORTADILLO:	1	0.000062	0.000062	0.999752	0.000061	0.000062
THE DECEITFUL MARRIAGE	1	0.000238	0.999049	0.000239	0.000234	0.000239
THE FORCE OF BLOOD.	1	0.000151	0.110696	0.000151	0.000148	0.888853
THE GENEROUS LOVER.	1	0.000071	0.000071	0.000071	0.000070	0.999716
THE ILLUSTRIOUS SCULLERY-MAID.	1	0.999760	0.000060	0.000060	0.000059	0.000060
THE JEALOUS ESTRAMADURAN.	1	0.000081	0.000081	0.000081	0.000080	0.999677
THE LADY CORNELIA.	1	0.000073	0.999708	0.000073	0.000072	0.000074
THE LICENTIATE VIDRIERA; OR, DOCTOR GLASS-CASE.	1	0.999577	0.000106	0.000106	0.000104	0.000107
THE LITTLE GIPSY GIRL.	1	0.000047	0.999811	0.000047	0.000047	0.000048
THE SPANISH-ENGLISH LADY.	1	0.000075	0.000075	0.000075	0.000074	0.999701
THE TWO DAMSELS.	1	0.000091	0.000091	0.000090	0.000089	0.999639

Figure 7. Title and Topic Associations in English

These heat maps indicate that in addition to having variability in the topics pulled from the text, the associations between the stories and their most identifying topic vary as well. In the associations generated from the Spanish text, LC is the only story most heavily associated with topic 1. Yet, in the English associations, LC is grouped with LG and DM. A more surprising observation becomes apparent when looking at the stories associated with the most prevalent topic from each language (with 9/12 most prominent words overlapping). The Spanish texts are grouped such that GL, DM, LV, SL, and LG are together and the English texts are grouped such that TD, SL, JE, GL, and FB are together. Although these two topics had a majority of the same words describing them, the associated stories are different.

Last, sentiment analysis was used to evaluate the overarching emotions associated with each story as well as the polarity. As mentioned previously, understanding the emotional similarities and differences induced by translation was a driving factor of this project and can be visualized in Figures 8 and 9.

title	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
COLOQUIO QUE PASÓ ENTRE CIPION Y BERGANZA	0.238743	0.291419	0.207307	0.335599	0.365336	0.292268	0.157179	0.395922
EL AMANTE LIBERAL	0.263277	0.284746	0.174011	0.324294	0.385311	0.319774	0.167232	0.402260
EL CASAMIENTO ENGAÑOSO	0.237864	0.320388	0.160194	0.300971	0.378641	0.310680	0.179612	0.490291
EL CELOSO ESTREMEÑO	0.203927	0.312689	0.125378	0.305136	0.364048	0.335347	0.164653	0.385196
EL LICENCIADO VIDRIERA	0.292740	0.306792	0.210773	0.358314	0.323185	0.297424	0.154567	0.381733
LA ESPAÑOLA INGLESA	0.209559	0.311275	0.126225	0.305147	0.388480	0.237745	0.162990	0.490196
LA FUERZA DE LA SANGRE	0.171429	0.319048	0.135714	0.292857	0.411905	0.321429	0.135714	0.473810
LA ILUSTRE FREGONA	0.219361	0.287333	0.187436	0.293512	0.380021	0.251287	0.156540	0.396498
LA JITANILLA	0.195015	0.412023	0.142962	0.282258	0.510997	0.229472	0.303519	0.369501
LA SEÑORA CORNELIA	0.175109	0.314038	0.159190	0.332851	0.363242	0.234443	0.143271	0.481910
LAS DOS DONCELLAS	0.228612	0.339411	0.150070	0.316971	0.363254	0.259467	0.166900	0.453015
RINCONETE Y CORTADILLO	0.264516	0.330645	0.164516	0.306452	0.369355	0.275806	0.220968	0.480645

Figure 8. Emotional Associations Spanish

title	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
DIALOGUE BETWEEN SCIPIO AND BERGANZA	0.282392	0.291528	0.211794	0.321429	0.352990	0.302326	0.184385	0.463455
RINCONETE AND CORTADILLO	0.226434	0.324795	0.141393	0.296107	0.375000	0.256148	0.160861	0.493852
THE DECEITFUL MARRIAGE	0.230769	0.303167	0.171946	0.212670	0.371041	0.257919	0.153846	0.529412
THE FORCE OF BLOOD	0.159910	0.315315	0.126126	0.272523	0.421171	0.299550	0.173423	0.463964
THE GENEROUS LOVER	0.271645	0.300866	0.156926	0.312771	0.380952	0.294372	0.189394	0.382035
THE ILLUSTRIOUS SCULLERY-MAID	0.200627	0.301985	0.121212	0.245559	0.422153	0.234065	0.175549	0.459770
THE JEALOUS ESTRAMADURAN	0.203947	0.317105	0.146053	0.259211	0.396053	0.276316	0.165789	0.435526
THE LADY CORNELIA	0.154525	0.306843	0.088300	0.273731	0.406181	0.207506	0.183223	0.464680
THE LICENTIATE VIDRIERA; OR, DOCTOR GLASS-CASE	0.242026	0.240150	0.185741	0.294559	0.313321	0.257036	0.114447	0.450281
THE LITTLE GIPSY GIRL	0.235507	0.343478	0.137681	0.294203	0.464493	0.256522	0.216667	0.457971
THE SPANISH-ENGLISH LADY	0.189130	0.319565	0.111957	0.260870	0.419565	0.256522	0.178261	0.498913
THE TWO DAMSELS	0.244565	0.316576	0.118207	0.305707	0.381793	0.250000	0.163043	0.414402

Figure 9. Emotional Associations English

Looking at the general trends in the emotions, the English translation appears to be fairly well in tune with the most significant emotions present in the stories, with the exception of LG. In both languages, trust, joy and anticipation play a big role in many of the stories. That being said, it is clear that there is still more variability and a wider range of observed emotions in the Spanish text as heavier weights are given to “fear” and “sadness”, and lower weights are given to “trust”.

Conclusion

Overall, this analysis has indicated that while the English translation of Cervantes' short stories was able to capture the primary emotions, topics and relationships in and between the stories, it fell short with some of the more nuanced sentiments. It also indicated that PCA is not very well suited for making comparisons between two corpuses unless they are being evaluated as one set rather than individually compared. In the future, it would be beneficial to take the sentiment analysis further and it may be interesting to chunk by paragraph rather than the story as a whole. Additionally, it would provide more depth to the analysis if other translations, both in languages similar to the original language and different, were included.

References

"Miguel De Cervantes." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., https://www.britannica.com/biography/Miguel-de-Cervantes.

"Middle Ages, Renaissance, and After." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., https://www.britannica.com/art/short-story/Middle-Ages-Renaissance-and-after#ref504324.