Data Science Project Report

Pattern recognition of reported speech in 17th-century diplomatic correspondence

Emma Zijp, Juan Paas, Sandro Kvrgić

Radboud Universiteit

Supervised by Gergana Slaveykova

13 juni 2025

1. Introduction

This project aims to systematically extract instances of reported speech from the correspondence of Christofforo Suriano, who served as the Venetian envoy to The Hague in the Dutch Republic between 1616 and 1623. During this time, the Dutch Republic was at war with the Spanish Empire to secure its independence, which made this posting an important one for the Venetian Republic's political and economic interests.
His letters, kept in the *Archivio di Stato di Venezia,* offer a rich window into the ways early modern diplomacy was conducted. Over the course of his time in The Hague, Suriano wrote 725 letters to the Venetian Doge - the head of state elected by a council of prominent citizens. They contain, among other things, detailed accounts of his interactions with important political figures in The Hague.

1.1. Historical context

Diplomacy has been an important part of international politics for centuries. International diplomacy and communication are essential for forming effective alliances and fighting wars. Currently, diplomacy might look like one leader making a phone-call to another, but in the 17th century diplomacy was conducted very differently. Diplomacy was based much more on audiences with leaders and envoys, due to the technological restraints on communication of the time. This diplomacy was based much more on face-to-face negotiation and personal interactions. In order to keep a written record of these encounters, envoys would send regular correspondence to their home countries containing large amounts of 'reported speech', sections which recorded what was said during the encounter (de Vivo, 2016).

These sections of reported speech were often some of the most important sections in this diplomatic correspondence. They formed the records of audiences with the political leaders which were at the core of diplomatic relations. Being able to study these interactions systematically means getting at the core of 17th century diplomacy (Van Faassen & Lamal, n.d.)

1.2. Linguistic context

Direct speech refers to a speaker repeating the original words used by themself or by someone else, meaning an exact quotation. Indirect speech, on the other hand, refers to a speaker adapting the direct speech by changing some aspects of the quote, such as the wording, pronouns or adverbs (Deng, 2023). Direct speech can usually be recognised easily by the use of quotation marks. In historical texts, however, these quotation marks are often absent or used inconsistently. It is therefore necessary to look at other linguistic markers or extralinguistic factors that might indicate direct speech. Reporting verbs are the most salient linguistic markers of direct speech. These verbs can be either in the present or in the past tense (e.g. "What she *said* was.." and "They usually *say*..") (Deng, 2023).

It is essential to establish a set of linguistic markers other than quotation marks indicating direct speech and investigate which ones are the most common. This is precisely

what Schlör et al. (2016) did. They found that in literary texts, long hyphens are the strongest indicators of the beginning of a quote, whereas the introduction of a speaker usually indicates the end (e.g. "…she said quietly."). Other frequently occurring indicators are reporting verbs (e.g. 'said', 'asked', 'answered'), question and exclamation marks, temporal adverbs (referring to time and place), and sentence length. These markers, however, were only found to be important indicators of direct speech when used in combination. Likewise, Van Faassen and Lamal (n.d.) propose that words that indicate an interaction (e.g. reporting verbs) in combination with proper nouns are an effective way of detecting direct speech. They also pinpoint a few extralinguistic features accompanying interactions that are useful indicators, such as smiling or shrugging shoulders.

Byszuk et al. (2020) constructed a model that identifies direct speech in 19th-century fictional prose that showed a performance of approximately 90% correctly identified direct speech. Major issues they faced were the linguistic variation often seen in historical context as well as obsolete vocabulary. There are a few prominent causes of errors the researchers give. The first is that the model also identifies first person narration as direct speech, which is usually not desired. The size of the paragraph also has a substantial influence on the recognition performance, with one-sentence paragraphs being identified as direct speech more often than longer paragraphs. Finally, the model also seems to rely on certain specific linguistic features, such as the imperative mood, proper nouns and interjections.

The aforementioned studies in combination with the historical context provide a steady base for our research. The aim of this study is to build an artificial intelligence (AI) model that can extract instances of direct speech in Christofforo Suriano's letters in the 17th century by looking at all sorts of linguistic markers as well taking the diplomatic context into account. Our research question is as follows:

How can we use an artificial intelligence model to extract instances of reported speech from 17th-century Italian letters?

2. Methods

2.1. Data

The dataset consists of transcriptions of 725 letters written by Suriano in 17th-century Italian, spread across ten Word documents and amounting to nearly 2,800 pages. Due to the lack of standardized spelling and punctuation conventions - particularly the use of quotation marks - automatically detecting instances of reported speech poses a challenge. Building on Lamal's (year?) study of these letters, Van Faasen and Lamal (n.d.) recommend a few approaches for automatic detection.

Firstly, they recommend looking for these other linguistic markers to indicate reported speech, which are reporting verbs combined with names of third persons as well as 'bodily expressions'. Using only the verbs as these linguistic markers would be less accurate in early modern correspondence. It is more beneficial to examine whole sentence structures, including

names, references to the third-person, and bodily expressions, which Suriano also used to illustrate the tone and emotional state of the other person in the interaction.

Secondly, they mention that parts of the letters were encoded with a cipher, in order to conceal the most important parts of the letters should someone intercept them. These parts are in italics, which is explained by Van Faasen and Lamal (n.d.) who write that “ciphered parts of the letter were very often related to relaying information about conversations and direct speech. (...) And it may be a potential route to detect speech passages using digital tools.” In order to start looking for patterns and linguistic markers to identify reported speech in early modern Italian and Suriano’s personal style, we built on these recommendations. It would be inefficient to start reading the letters from the beginning to the end. We therefore decided to focus on the ciphered parts of the letters, as they “were very often related to relaying information about conversations and direct speech”. We concluded that these parts would have a greater chance of containing reported speech than the rest of the letters.

Looking only at the italics in the word document, we started to go through it by hand. We searched for any linguistic pattern to mark the beginning, middle and end of reported speech. It is important to not just find a marker to indicate the beginning, but when detecting it automatically, there needs to be something marking the end. We were looking for a pattern in verbs, structure with names or indicators of a third person, as well as bodily expressions. Some examples include:

- “et a me lo diceva in confidenza” -> he said to me confidentially
- “suo Principe mi raccordava” -> his Prince advised me that
- “mi rispose con faccia allegra” -> he replied to me with a cheerful face
- “a me ha detto liberamente con queste formali parole” -> has said to me openly/freely with these formal words
- “et continuò” -> and (he/she) continued

Identifying the beginning of reported speech is relatively simple, as it is often preceded by a reporting verb (‘mi rispose’ - he said to me; ‘Disse’ - he said). The ending of the reported speech, however, is vastly more difficult since there is a lot of variation. We had to look for certain patterns in the endings, though sometimes there was no clear ending whatsoever. The most common indicator is when Suriano responds to the person who was just quoted, meaning that Suriano gives an account of what this person said (‘Io disse’ - I said; ‘Io risposi’ - I responded; ‘Io aggiunsi’ - I added). Sometimes Suriano is giving information in the form of a dialogue and then proceeds to give his opinion on the just reported speech. This is a somewhat common indicator of reported speech ending. There are also some constructions of report speech endings that occurred very rarely, which include:

- “l’ho dato reverente conto” -> I have given it reverent account (He has honourably recited the speaker)
- “Il profferì con qualche sorriso” -> He said it with a smile
- “il che mi è sta[to] detto anco da altri” -> which I was also told by others
- “vedendo ch’io a ciò non rispondevo” -> seeing that I did not respond to this

An issue in detecting direct speech is that it is not always possible to make a clear distinction between direct and indirect speech. It can be difficult to say whether Suriano was

paraphrasing or quoting someone. We were therefore quite strict when manually determining what was a marker of reported speech.

2.2. Building the model

The goal of this project was to find a way to extract reported speech from the letters by using Artificial Intelligence techniques. We worked in the programming language Python, using Jupyter notebooks and csv files. The technical part of the project was started by loading the official text-fabric corpus of the project. Given our lack of knowledge with this kind of data, it took us some time to load the text part of the data. Our knowledge about the ciphered text in the italic data moved our attention to trying to build our model on the italic data, and subsequently applying that model on the whole data. We therefore started with the extraction of italic data from the corpus, which did not yield any results as the corpus did not have any italic data or markers for the italicised data. This meant that we would have to use our local dataset made from multiple word documents, as that dataset contained italic texts. We then extracted all of the italic parts from all ten documents and started to string match the italic dataset from local documents to the corpus dataset. This essentially allowed us to have italic sentences from the corpus dataset. We encountered many problems while doing this and in the end realised that it was unfeasible due to many formatting and alignment issues. This led us to abandon the corpus dataset and focus only on the italics dataset we just created.

The next phase of the project started with our work on the italics dataset. As the data we extracted was not sentences, but plain text, we built a sentence splitter. We had to correct this part of the code a few times during our project, after manually inspecting the output document, as we could not reach the needed consistency with it. There was also a problem with the full stops inside the italic text, which were not all italicised. We therefore manually inspected and edited every document, as to not miss any edge cases. We changed the format of full stops to italics, making sure that we split each sentence correctly. This led to an additional 70 sentences on top of our 630. We also encountered a problem of Italian special letters not being encoded properly into our csv files, which instead showed up as ambiguous signs. We therefore needed to add additional preprocessing steps, using python library ftfy, which ensured the dataset was clean and ready for use.

We then continued by building a complete regex system for finding reported speech patterns. We collected all of the knowledge about patterns in reported speech by manually analyzing the letters and subsequently applying these patterns. This system is the core of our project, and is of great significance in other systems which were later built around it. One of the systems built around it was the automatic labeling of direct speech sentences. This enforced the regex patterns, as well as having some additional rules within the system's functionality, ensuring the maximization of our task with the knowledge at hand. This system was not only used to label reported speech, but it was also used to add additional labels such as previous_reported and next_reported. The latter two had a boolean output for each sentence, allowing us a more context-based labeling. This allowed us to put additional attention on sentences around reported speech, which have a higher chance of being reported speech. This system worked by giving a positive boolean value for each sentence where the next sentence was labeled as reported speech, and at least one of the two previous sentences

was labeled as reported speech as well. This method allowed us to add additional features on which our machine learning model would be trained, adding more robustness and possibly better performance.

The performance of the automatic labeling system exceeded our initial expectations. Nevertheless, we decided that the best course of action would be to train our machine learning model on 100% correct labels. We therefore labeled all of the data by hand, making sure to fix the system's labeling mistakes, and then used preprocessing methods to adjust labels previous_reported and next_reported based on our new ground-truth annotations. After we had our data correctly labeled, it was time to start building the model.

We decided to go for a relatively simple but effective solution, combining logistic regression classifier with Term Frequency - Inverse Document Frequency (TF-IDF) Vectorization. TF-IDF transforms text into numerical features based on two principles: term frequency (TF), which reflects how often a word appears in a given sentence, and inverse document frequency (IDF), which reduces the importance of words that are common across the entire dataset. This results in a vectorized representation of each sentence, which can be fed into a machine learning model. We choose logistic regression for its binary property, allowing the model to assign a boolean label to each sentence. Its simplicity, interpretability and speed made it very effective for our task. After building the initial model and subsequently training it, we wanted to ensure that its performance was not the result of overfitting or specific to a single train-test split. We introduced 5 fold cross-validation across the entire labeled dataset in order to evaluate its generalization ability more robustly. During each fold, data was split into 80% train and 20% validation, leading to some interesting results in the end.

For the final part of this project, our goal was to generate a tangible and useful output for the project owner. We decided to use our automatic regex labeling system and apply it to all of the 725 letters. This allowed us to produce two csv documents, with one containing all of the sentences from letters labeled, and the other only containing reported speech from the whole dataset. Due to time constraints, we were unable to clean up the whole dataset from metadata, as well as label it ourselves. This would enable us to make a new machine learning model which would be trained on a bigger dataset. Nonetheless, the document containing only reported speech provides a highly useful output for the project owner, especially for historical or linguistic analyses.

3. Results

The first method we are going to discuss is the rule-based regex labeling system. This method produced excellent results on our italic dataset. It managed to achieve an accuracy of 95%, which we calculated by manually checking the predicted labels for each sentence. The real percentage might be slightly lower, as some cases were ambiguous as to whether they were reported speech or not. For these cases, we deferred to the predicted label. It is important to note that such ambiguous cases were rare and are unlikely to have significantly affected the overall accuracy.

The second method, the machine learning model combining TF-IDF with the logistic regression classifier, also produced great results on italic data. After 5 fold cross-validation, it had an average accuracy of 79.1%, with precision of 62.7%, recall of 69.8% and macro F1 score of 65.9%. These results indicate that the model is able to find reported speech patterns. Further evaluation on a hold-out test set, which was made from 20% of the data, resulted in an even better accuracy of 81%. These results prove that even a relatively simple model, when combined with meaningful features and high quality data, can perform effectively in complex linguistic tasks.

Metric	Cross-Validation	Test set (Overall)	Test set (class 0)	Test set (class 1)
Accuracy	0.791	0.81	-	-
Precision	0.627	0.86	0.86	0.67
Recall	0.698	0.81	0.87	0.65
F1 score	0.659	0.81	0.86	0.66
Support	-	139	99	40

Figure 1: This is the table representing all of the results from the model applied on the italic data. It contains both Cross-Validation and hold out test results. It also presents results for each class in the dataset, where class 0 corresponds to non-reported speech and class 1 to reported speech.

4. Discussion

4.1. Implications

Being able to systematically extract instances of reported speech from the letters of Suriano, means being able to get at the heart of diplomatic relationships between the Dutch and Venetian Republics of the early 17th century. His correspondence is a great source of insight into Dutch, Venetian and European history at a turbulent time of war on the continent. The Dutch were fighting against the Spanish in their Eighty Years' war, while at the same time, the rest of Europe was plunged into the series of wars we now call the Thirty Years' war. The reported speech in his letters detail the most important pieces of diplomatic interactions, and by extracting these pieces of text, this project opens up another way into the study of the political forces behind these destructive wars which shaped the next century of European history.

It also helps to study the reported speech extracted here on its own merits. They show how diplomacy was conducted, what could be said and what information could be passed on to the head of state in the home country of the envoy. This can tell us something about the flow of information and news in the 17th century.
Even with imperfect results or the limitations of this project, the work towards historical research with the results accrued here can continue. If this model could be developed further

and applied on more early modern Italian correspondence, even imperfect results could mean new ways of studying these historical interactions.

4.2. Limitations

The most crucial limitation of this study is that we had to label the data manually. First of all, this was a time-consuming task, limiting our time we could spend on building the model. The main problem, however, is that none of us speak Italian, let alone 17th-century Italian. We therefore had to rely on online dictionaries and translation machines, which only know modern Italian. This made it significantly more difficult to understand the text in a language that we are not familiar with. Identifying direct speech in the letters was thus a complex task and as our results were based on the labeled data, the reliability of the final results is diminished. Another problem that occurred as we were manually labelling the data was that the italicised texts were sometimes ambiguous as to whether it was actually reported speech or not. The lack of standardisation of 17th-century languages, in combination with our lack of knowledge of Italian, makes it difficult to distinguish direct from indirect speech.

Another limitation of our study is that our model is only trained on instances of direct speech in the italicised text rather than the full dataset. Our partner informed us that the italicised texts contain ciphered messages since they contain highly confidential and personal information. These ciphered messages most likely contain instances of direct speech because they are Suriano's accounts of what important officials reported to him confidentially. We therefore decided to solely focus on the italicised text, also because of time constraints. It is thus unknown how the model would work if the entire dataset were labelled for direct speech and if the model was subsequently trained on this extensive dataset.

4.3. Future research

We propose a few improvements future studies can make to build a more reliable model that can detect direct speech. First of all, this type of study would benefit greatly from having a researcher with knowledge on Italian and especially early modern Italian. This would make it significantly easier to label the data, making the results more reliable. Another improvement a future study could make is applying this model on the full dataset rather than merely on the italic parts. Each one of Suriano's letters would first have to be labelled for direct speech, which was impossible for us given the limited time. The model would then be trained on the entire dataset, creating once again more reliable results. Lastly, to ensure the model is accurate enough, it can be applied on an official corpus of early modern Italian texts. This can verify how well the model works.

5. Conclusion

This study aimed to investigate the feasibility of using an AI model to extract instances of reported speech in 17th-century diplomatic Italian letters. We have been able to demonstrate that meaningful linguistic patterns can be extracted from historical texts using artificial intelligence. We manually labelled the italicised texts in Christofforo Suriano's letters and trained our model on this data. Our rule-based regex labeling system, which was based on this manually labelled data, showed highly accurate results. Our machine learning model likewise showed accurate results. Despite methodological challenges and time constraints, this study lays the foundation for future research on detecting direct speech using machine learning.

6. Bibliography

De Vivo, F. (2016). Archives of speech: Recording diplomatic negotiation in late medieval and early modern Italy. *European History Quarterly*, *46*(3), 519–544. https://doi.org/10.1177/0265691416648275.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U., & Calvo Tello, J. (2016). Straight talk! Automatic recognition of direct speech in nineteenth-century French novels. In *Digital Humanities 2016: Conference abstracts*, 346–353. Jagiellonian University & Pedagogical University, Kraków.

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., & Eder, M. (2020). *Detecting direct speech in multilingual collection of 19th-century novels*. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, 100–104. European Language Resources Association (ELRA). https://aclanthology.org/2020.lt4hala-1.15/.

Van Faassen & Lamal (n.d.). *From formulaic to forms? Tracing the evolution of speech and information flows in diplomatic letters*. [Unpublished manuscript].

Deng, D. (2023). "She's like why you speak English while dreaming?": A corpus-based study of quotative markers used by Chinese speakers of L2 English. *Languages*, *8*(1), 51. https://doi.org/10.3390/languages8010051.