ENCRYPTED EPIGRAPHS

A DECRYPTION REPORT OF THE MYSTERIOUS EPIGRAPH IN THE NEAPOLITAN CHURCH OF SANTA MARIA LA NOVA

Cosimo Palma, Independent Scholar
cosimo.palma@protonmail.com

ABSTRACT

This paper contains all steps regarded as necessary and relevant to achieve the decryption of the epigraphy[1] placed in the Turbolo Chapel within the Neapolitan Church of Santa Maria La Nova. The epigraphy has been processed by means of a Python- based algorithm, which helped to point out and display linguistic features converging univocally to the encryption hypothesis: the epigraph hence has been written in a real language and encrypted by simple characters substitution. A preliminary analysis on lexical diversity, diphthongs' and triphthongs' count, vocals' and consonants' intertwinement, palindromes and anagrams have strongly pointed to Coptic, although Greek and Latin still preserve some peculiarities preventing us from discarding them altogether.

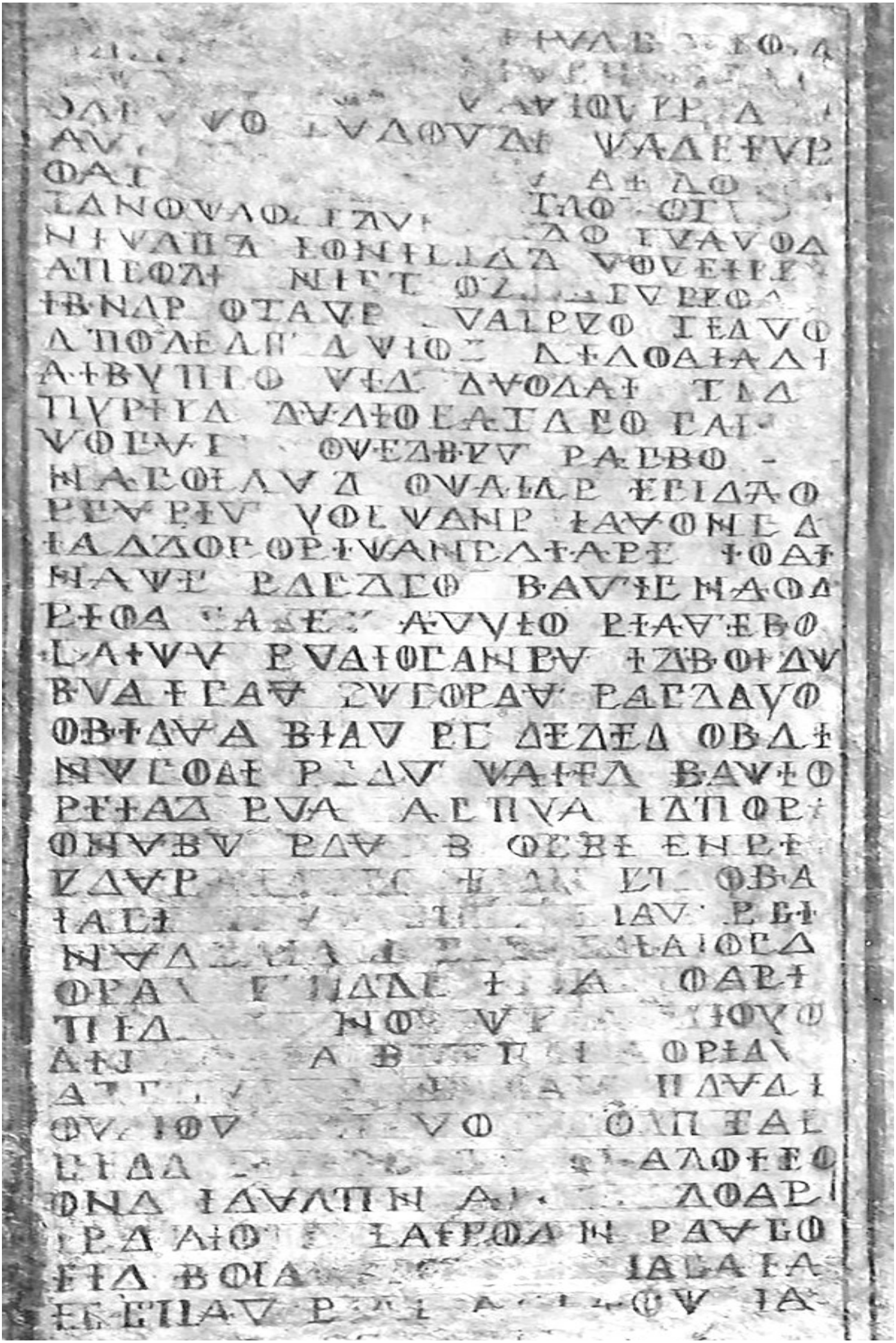
INTRODUCTION

Within the Turbolo Chapel of the Santa Maria la Nova church in Naples (Italy) two epigraphies are to be found. The one on the observer's left side contains an indulgency statement from the Pope Gregory XIII related to holy masses celebrated therein, in dedication to Miss Turbolo, the noble woman Giovanna De Rosa. On the right side we find an epigraph written in an *unknown* language (Pic. 1). This immediately recognisable feature does not let us understand that we behold an encrypted text yet: as often happens in these cases, the only decisive proof we are coping with an encryption would be to find out its decryption, or at least any other external element which could let us pointing to it as such without further doubts.
Although there is no absolute certainty towards this direction, a queue of observations has compelled me to try out the decryption path:

- The amount of different glyphs matches pretty well the amount of letters within the Greek, Coptic and Latin alphabets[2];
- The glyphs' frequency distribution of the encrypted epigraph reflects the character's frequency distribution of Latin-,Greek- and Coptic language corpora well enough to lead us to the assumption we are dealing with a natural language;
- The character's distribution of a random text (with randomly generated letters, not words!) looks on the other side completely different: the frequency range is far smaller, the function behavior better modeled by a regression line, presenting many *plateaux*;
- The distribution of the types' length reflects pretty well the same occurring in natural languages;
- The variable distance of the board's right margin of the epigraphy lets us understand that it is never the case that a word is cut at the edge and then continued on the following line. This observation together with the clear decreasing size of the letter from the left to the right let understand that the writer tried to fit into the given space, a possibility which would make no sense in the scenario where all words were randomly written.

1 To become aware of all the steps, even the wrong or superfluous ones, through which the whole research unravelled, please refer to "Encrypted Epigraphs II - a research journal about the mysterious epigraph in the Neapolitan church of Santa Maria la Nova", which nonetheless must not be regarded as a mere collection of failed attempts and impasses, rather as a *vademecum* for the approach to all sorts of similar problems (i.e. Encrypted Epigraphy).

2 The amount of glyphs seems to be actually slightly inferior or superior, depending on the language we compare to. It is not uncommon in cyphers that the relationship between clear text- and cipher text-characters is not perfectly bijective.



Picture 1: The Encrypted Epigraphy of the Turbolo Chapel in the Church of Santa Maria La Nova

Hence my choice to label the analysed artifact as “encrypted” and not as “mysterious” or “secret” appears to be justified by reasonable grounds, rather than only by a scientifically sound wish for exactness.

LETTER-GLYPH MAPPING

Below the letter-glyph mapping which was necessary establishing to produce a machine-readable document file so that the text could be analysed and processed by means of algorithms[3]:

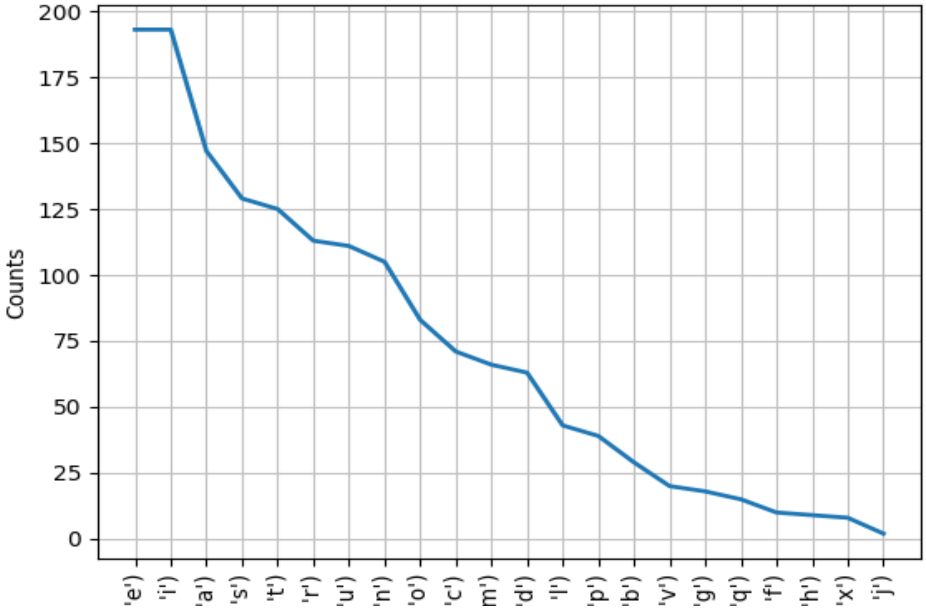
A >>> 
B >>> 
C >>> 
D >>> 
E >>> 
F >>> 
G >>> 
I >>> 
J >>> 
K >>> 
L >>> 
M >>> 
N >>> 
O >>> 
P >>> 
Q >>> 
R >>> 
S >>> 
T >>> 
U >>> 
W >>> 
X >>> 
V >>> 
Z >>> 

A GLYPH'S FREQUENCY DISTRIBUTION COMPARISON AMONG LATIN, GREEK AND COPTIC

The latin epigraphy contains many symbols standing for longer and frequent sillables, like for instance "9" as "us". The character distribution below takes into account the plain text, without such symbols and also inserting potentially missing letters for brevity's sake (like for example, "domini"

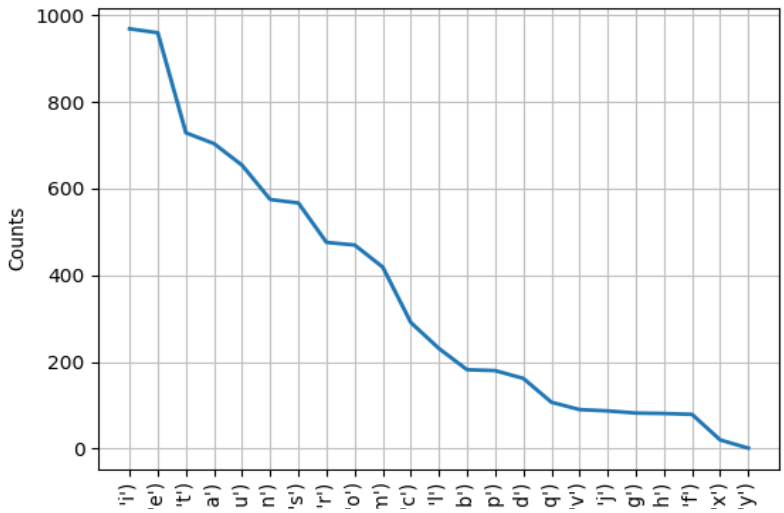
3 The whole Python project can be found at https://github.com/Glottocrisio/MariaLaNova

for "dni"). The very nature of this preventive analysis, which still is not tailored for exactness, let me ignore these small adjustments momentarily. A pure "letter frequency" approach for the whole decryption could be considered only in the case that the chosen language letter frequency distribution of the encrypted text's length is comparable to the corpus' one.[4]
For the very same reason at this stage any kind of mathematical formulation of distributions' divergences is avoided.



Plot 1: Letter frequency of latin epigraph from Pope Gregory XIII

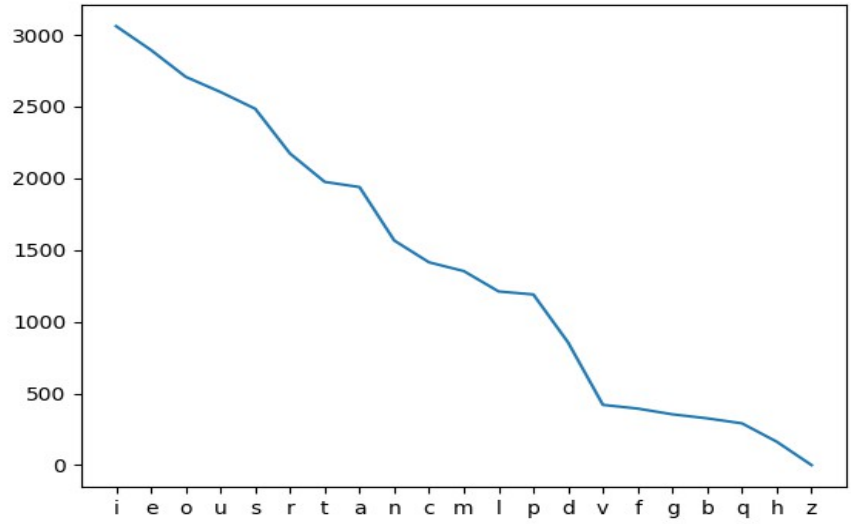
In Plot 1 and Plot 2 is possible to notice among others, that in a natural language text the letter frequency ranges between 0 and a number, which usually corresponds to about *one tenth of the overall characters amount.*
Below is how the letter frequency distribution looks like for the Universal Declaration of Human Rights in Latin[5]:



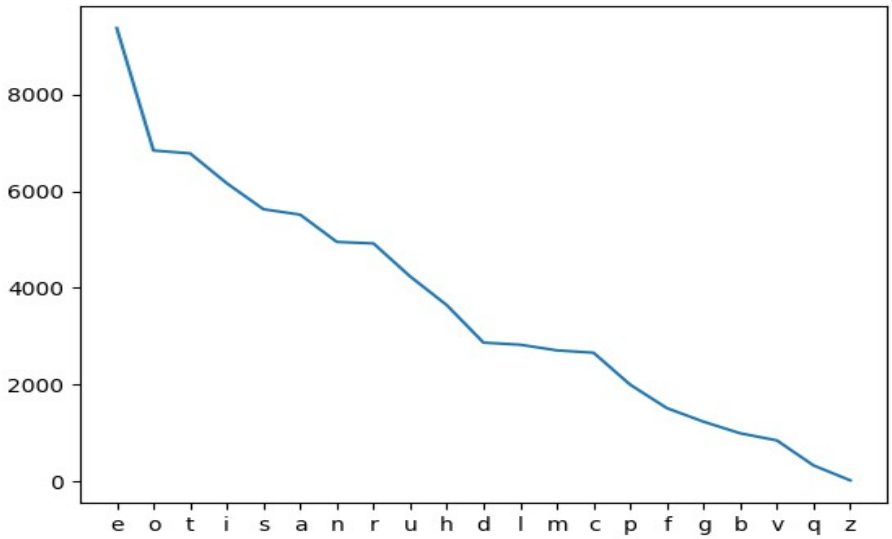
Plot 2: Letter frequency of Latin Universal Declaration of Human Rights

4 Even this extreme case may be valid only if we are almost certain of a biunivocal relation among clear text and encrypted characters, an eventuality I am not able to exclude yet.
5 Universal Declaration of Human Rights (UDHR) corpus

Plot 3: Letter frequency distribution of 1000 random sentences of the Perseus Latin Corpus

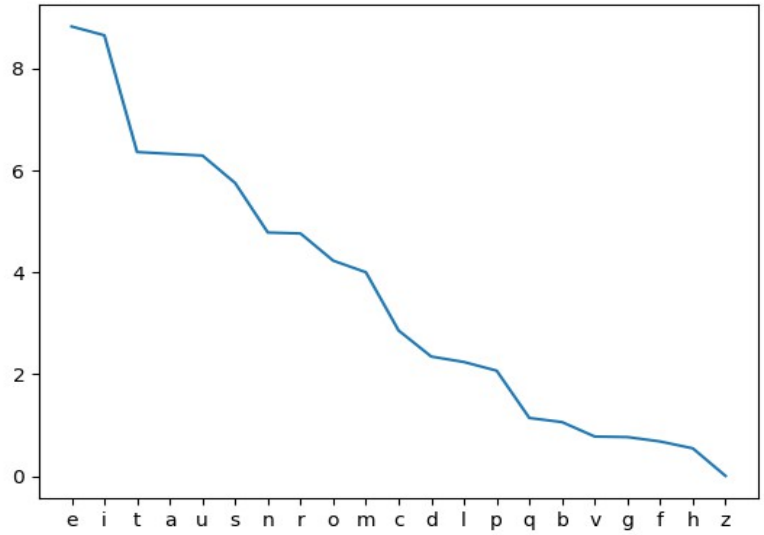


From a comparison between the latin epigraphy and the Universal Declaration of Human Rights it is already possible to recognise similarities, such as the same function behavior and almost the same letter ranking.
On the other hand, a broader sample, namely the *Perseus Text of Classic Latin* (Plot 3) shows a surprising feature: although the glyphs' sequence on the x-axes does not significantly vary, the function's shape is completely different, a fact which could maybe be justified but the differences between the modern/ecclesiastical Latin of the papal inscription and the ancient/classical Latin of the Perseus Corpus. The letter distribution for one thousand sentences is very similar to the one for only fourty sentences, in the context of the same corpus (plot omitted for brevity's sake).



Plot 4: Letter frequency distribution of whole Perseus Latin Corpus

The same distribution for the whole latin corpus is indeed far closer to the observed one for ancient Greek, and shows a pretty different letters rank, as well as a different function behavior.
To prove the diachronic difference among classical latin and and medieval latin the same letter count has been performed o a broader corpus, called „Latin Text Library“, including more than the double words' amount than the Perseus Corpus.

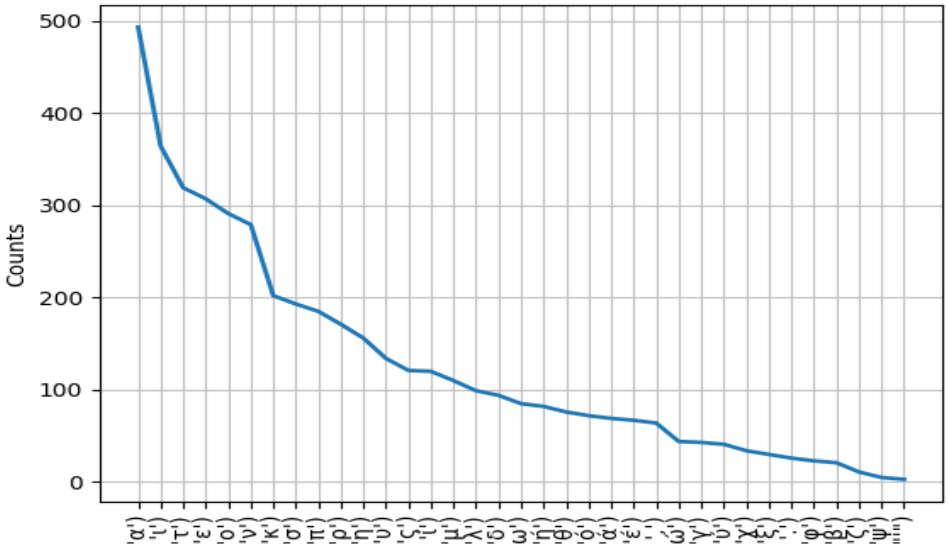


Plot 5: Letter frequency distribution of Latin Text Library

The distribution curve has shifted again to the previous shape, as well as the letter sequence on the x-axis. Thus, *the thesis of the different curves for the same language can be addressed indeed to the diachronic shift.*

This feature let us at least momentarily discard the eventuality we are dealing with medieval Latin (pretty much overlapping with the ecclesiastic one), because as we can notice from the glyphs' distribution of the encrypted epigraphy (see Plot 7 in the next page).

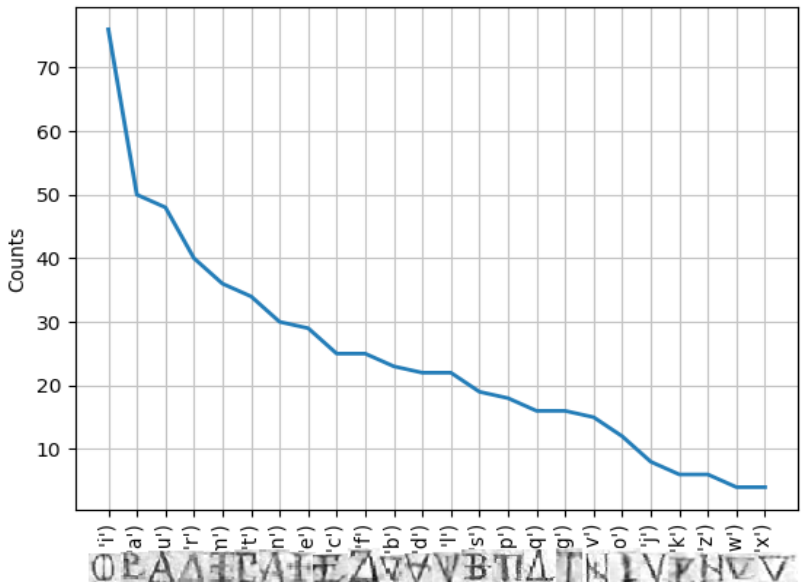
We are now esamining the letter frequency distribution fashion for the Greek language Corpus.



Plot 6: Letter frequency distribution in the New Greek Universal Declaration of Human Rights

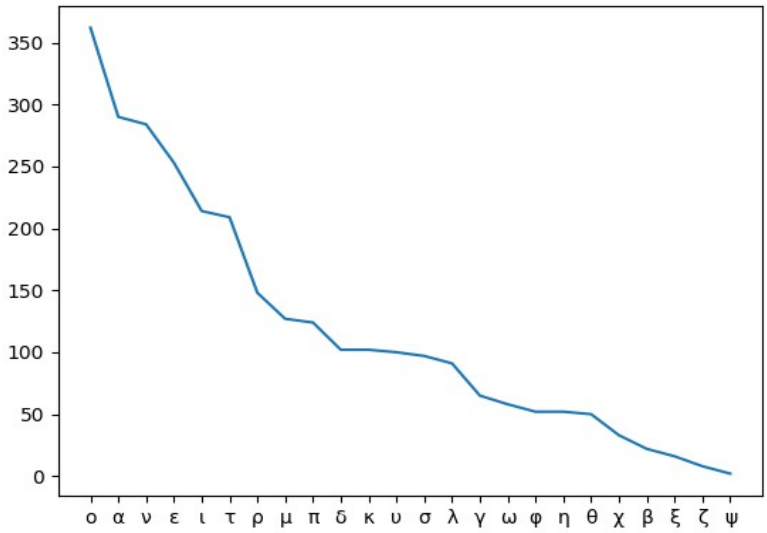
The glyphs frequency distribution of the New Greek language has been taken into account, because probably closer to the Greek, which eventually would have been used to encrypt the mysterious epigraph (of course excluding the case that the epigraph is a quote of classic Greek literature).

As a matter of fact, also the *koiné* Greek, through which the Gospels have been edited, is more similar to the modern Greek than the ancient.



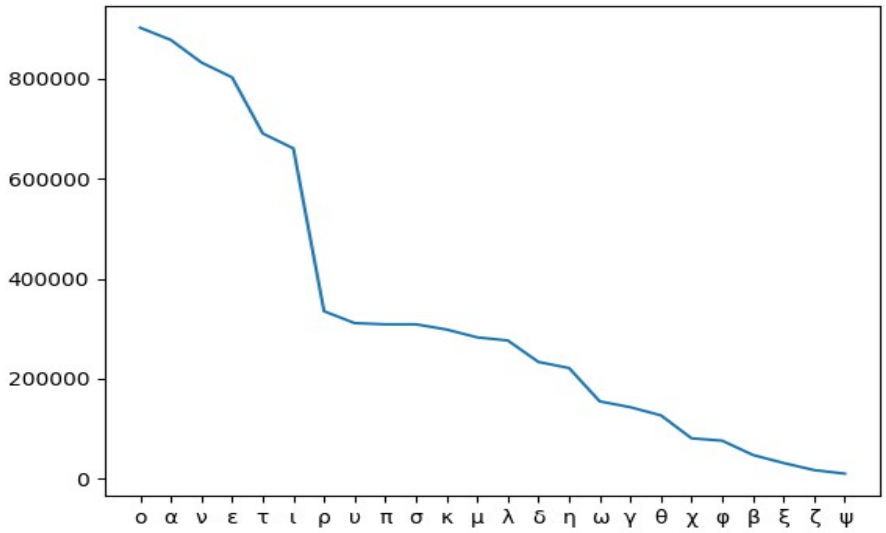
Plot 7: Letter frequency of mysterious epigraph

The glyphs' distribution of the mysterious epigraphy resembles in many details the Letter frequency distribution of 40 random sentences in the Perseus Ancient Greek Corpus and as well the Letter frequency distribution of the whole Perseus Latin Corpus and the Letter frequency distribution in the New Greek Universal Declaration of Human Rights. For this reason it seems to be impossible to exclude a language *a* priori. Besides the letter frequency distributions, it must be noticed that the *amount* of glyphs reflects perfectly the Greek language one. The indeniable resemblance of the glyph with the Greek alphabet and the observation performed in the following paragraph let us opt for this language to begin with, although, on the other hand, the consistency with other factors and elements external to the epigraph let us not neglect the Latin option. The slightly superior tendency to the Greek, further consolidated through the considerable frequency difference between the first two ranked characters, which we have already stated before be almost equal in latin samples of short extension, further strengthens the path to the Greek language.
For completeness' sake, in the following plots we present a closer look also to the Ancient Greek letters distribution.



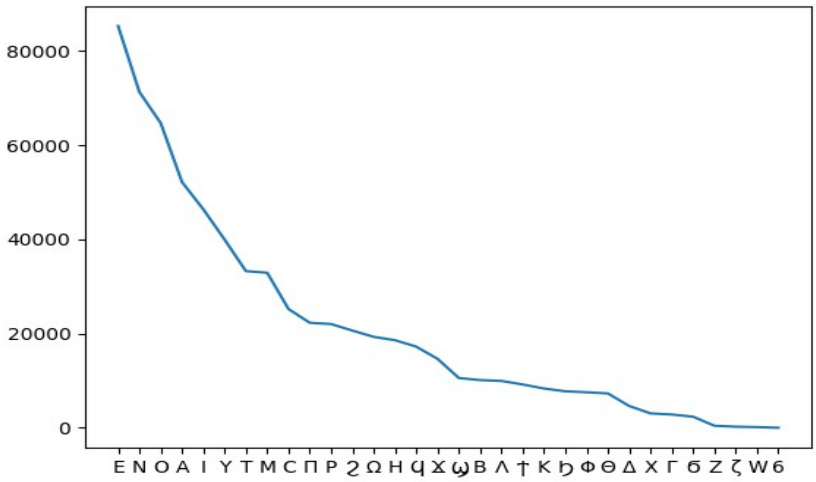
Plot 8: Letter frequency distribution of 40 random sentences in the Perseus Ancient Greek Corpus

Also for the ancient Greek we can observe the same weird behaviour of the Latin distribution on large samples: the curve for a fourty-sentences sample is pretty much identical to the one for 1000 sentences (also in this case the graph is omitted for brevity's sake)[6]. If we perform the frequency letter count on the whole corpus, the function's shape changes again drastically, a behavior that cannot be immediately explained otherwise than the Latin case.



Plot 9: Letter frequency distribution of whole Perseus Old Greek Corpus

The following plot has been modelled according to Coptic New Testament[7]. Also in this case the letter distribution seems to fairly mirror the epigraph's glyph's distribution, which was an unexpected behaviour, given the absence of parenthood between Coptic and Greek, despite the extreme similarity of their writing system.

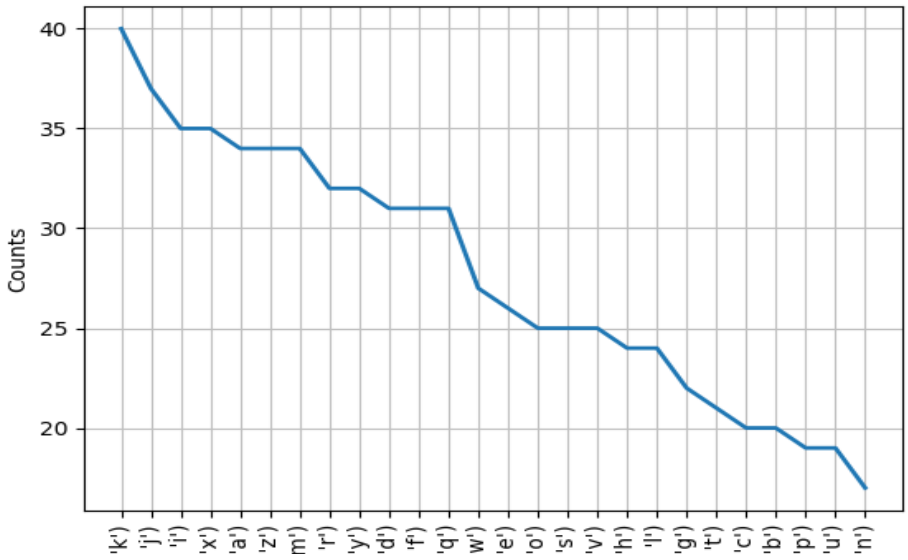


Plot 10: Consonant frequency distribution for the Coptic New Testament

If the previous plots have not been enough explicit to push our analysis in the "decryption" direction, the letter frequency distribution of a random text shows us with any further doubts, that we are on the right path.
As already pointed out at the beginning, the frequency range (for the same letters' amount) is far

6 Furthermore it is interesting to observe that the number of glyphs is in this case perfectly matching, whereas the general function's shape is maintained although the sequence of glyphs on the x- axes has changed.
7 Christos-c/bible-corpus

smaller, the function behavior better modeled by a regression line, presenting many plateaux.
For completeness' sake, in the next paragraph other observations will be listed, before proceeding in designing our solution for this interesting historical puzzle.



Plot 11: Letter frequency distribution of a random text

FURTHER OBSERVATIONS

- The comparison between both Latin and encrypted epigraphs does not show the latter to be a translation of the former[8];
- The epigraph's analyses based on fluorescence induced by ultraviolet radiation, on infrared rays and on pigment samples revealed that the dating of the inscription could also date back to the sixteenth century[9];
- The distribution of words' final letters is very skewed. Some of them even appear only in the second half;
- There are many glyphs which seems to be similar, excepted for some diacritic signs. It is to be understood if their impact is merely phonetic (same letter, different tone) or semantic (different letters). The total amount of glyphs is equal to the Greek/Coptic characters amount, and almost equal to the Latin one, a fact which lead me to opt for Greek[10];

ANALYSIS OF DIPHTONGS, TRIPHTHONGS, PREFIXES AND SUFFIXES

A function listing all diphthongs and triphthongs in the epigraph's string and in the Latin and Greek corpora cleared out my sight to further considerations. Usually, whether we are considering Latin or Greek, we can observe a suffixes' repetition which is typical of flexive languages. In the epigraphy this does not happen, or at least not as often as it should. This may lead only to four possibilities:

- either the author, conscious of this language feature problematic for an encryption, reallocated the suffixes somewhere else within the word, or even deleted them;
- or the used words are anagrams;
- or the language we behold is not a flexive but an agglutinative one (which in turn would

8 The deduction being based on words' and words' length counts, may be accountable only for languages belonging to the same category (flexive languages such as Greek and Latin). For agglutinative languages, such as Coptic or Hungarian this door shall be still considered open.

9 https://www.ilmattino.it/napoli/cultura/tomba_di_dracula_napoli_scoperta_misteriosa_iscrizione-4117802.html
This is the same datation of the latin epigraphy and of the whole Turbolo chapel.

10 The diacritics in the epigraphy may well be a clue about the character itself. For instance similar characters could be close in alphabetical order.

explain the presence of excessively long words, which cannot be found in any Greek and Latin corpus);
- The space between words is set randomly, and is not actually semantically cogent.

Because of all these factors, it is not only relevant to detect diphthongs and triphthongs frequency, but also to determine their position in the word.
As performing this task has shown (through a simple extension of the same function) the epigraphy shortness does not allow us to determine exactly which diphthongs and thriphtongs are more prone to figure as prefixes and suffixes.

THE ANAGRAM HYPOTHESIS

The type-token ratio[11] of the epygraph is 1, which may be a normal value for an extremely short text, but not for such one. The same value for an ancient Greek- and Latin- corpora 100-words sample is to be placed at around 0,8, while for Coptic at around 0,92. This particular feature could be also explained, for instance, in the case that the words are anagrammed.
The strategy I chose to detect anagrams was the alphabetical sorting of each word. Through this process apparently different words may have resulted to be identical. Pitifully also in this case the type-token ratio has shown the same result. This can mean different things: either the epigraph's author, aware of the fact that co-occurrences in an encrypted text constitute a flaw, carefully decided to avoid them, or there are indeed glyphs which may correspond to more letters. Another scenario, not likely but still possible, is that this case is totally casual, and maybe same occurrences of the visible words may be shown in the deleted parts (thus offering a plausible excuse for their deletion). The only two ways otherwise possible are:

- a high incidence of proper names, an eventuality which would neutralize all the advantages given by working with a corpus;
- White spaces to separate words are not set at the natural word end, but randomly, as also stated beforehand.

LONGEST WORD - SHORTEST WORD APPROACH

Afterwards I looked for the longest word in the epigraph, and compared to all words in the language-related corpus with same length. Principle of this choice, is that according to synergetic lynguistcs[12] the longer a word is, the rarer it is. This maximazes the probability of being able to actually match correctly words from the corpus to the epigraph, because of course the number of words having same length and the same letter sequence (including cooccurrences, as almost always happens with vowels) is sensibly smaller. In this way, we can determine at once the hugest amount of letters, which would speed up a lot our decryption process.
In the following function I compare a word from the epigraphy to all words of a given file (containing all words of equal length as the input word from a given language).
The output for *"outnmn"* will be: 123454, which I define as „superword“, a user friendlier version of a regular expression.

```
word = "outnmn"
matchWord(word, "coptic"+str(len(word))+".txt")
```

11 The *type-token ratio* is a parameter which returns a degree of lexical diversity in a given text by dividing the amount of different words (the set of words) by the amount of all words, disregarding if they occur more than once. The lexical variety has to be hold as higher as far as the TTR tends to 0.

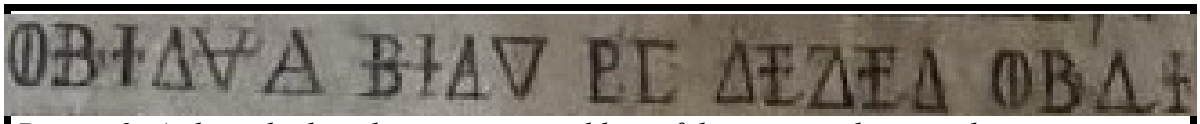
12 Köhler R. (1986)

```
def matchWord(word, file):
    m = open(file, "rt", encoding="utf-8")
    #output file to write the result to
    mw = open("matchWord.txt", "a",  encoding="utf-8")
    mwf = open("matchWordFile.txt", "a")
    m = m.read().split()
    superword=""
    superwordfile = ""
    superchar = ['1','2','3','4','5','6','7','8','9', 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'L']
    for char in word:
        superword = superword + superchar[list(word).index(char)]
    print(superword)
    #for each word in the input file
    for word in m:
        for char in word:
            superwordfile = superwordfile + superchar[list(word).index(char)]
        mwf.write(superwordfile + '\n')
        superwordfile = ""
    mwf = open("matchWordFile.txt", "rt")
    mwf = mwf.read().split()
    for word in mwf:
        if word == superword:
            mw.write(word + '\n')
    mw.close()
```

If there is one or more matches, we shall look for the *shortest* terms in the epigraphy containing as many glyphs as possible as the ones contained in the matching term. If after replacing all the original glyphs with the characters we behold a meainingful word in the language used for the comparison, we can assume with huge confidence that this very one is the source of the encryption, and slowly, by repeating the above-mentioned steps, all other letters can be successfully matched. Starting from this point it would be possible to work in a sort of *Constraint Programming (CP)* fashion, so that the consistency of every letter attribution may be addressed more consistently, being sure that all constraints adds up and are fullfilled at the same time. In our particular case this option does not seem very walkable given the lack of a „string-based“ Python library for CP, and given the very small size of our encrypted text.
Pitifully no 18-letters long Coptic word from the Coptic Bible has matched the correspondent epigraphy´s type. The only task that can be accomplished before moving forward, is trying with a bigger and more inclusive corpus, which is very difficult in the case of Coptic, a language whose documents still need to undergo huge digitalisation.
In the case no language gives positive results through this approach, it should be proceeded with a similar „looped comparison“, this time avoiding white spaces.
Of course the whole corpus will be taken in consideration then, and not only the list of words of equal length. It will be started as well with the longest epigraph's word, and the match will be tried starting from every position.

A CASE FOR GREEK: AN ANALOGIC ATTEMPT

The line 22 represented for me a good starting point for an analogic attempt, given the repetitions of letters taking place therein, as well as the presence of a 5-letters palindrome. A manual attempt with ancient Greek, has delivered the following result:



Picture 2: A closer look to the twenty-second line of the encrypted epigraphy

Α.Γ.Ο.Ρ.ΑΙ.Ος Κ.Ό.Ρ.Ος Δ.Έ Ρ.Η.Τ.Η.Ρ Ἀ.Β.Ρ.Ό

... vulgar servant, but a refined rethor...

This case has been the proverbial drop compelling me to the implementation of an algorithm which could speed up significantly this painstacking work of comparison and analysis.
My analogic mapping is in fact quite imaginative, since with no criteria whatsoever single glyphs are assigned to syllables (see **ΑΙ** and **Ος**), whereas diacritic variation are sometimes ignored and sometimes considered enough to pick up another letter (compare for instance the **Β**s in the first and in the last letter).

QUEST FOR PALINDROMES

There are no other palindromes in the whole epigraphy. Beside *ρητηρ, νομον* and *νοθον* seem to be the only two 5 letters palindrome to be found in the Ancient Greek Corpus. Since the latter two present glyphs that cannot be taken in consideration (*ν* and *ο*) because in a frequency range which differs too much from the one of the two original glyphs, assigning to these epigraph's glyphs the greek letter *ρ, η* and *τ* is the only solution contemplated withing the Greek language. At least for the moment, we can also put aside the Greek: the presence of only one matching palindrome in the whole corpus and other observations on the letter distribution let us opt for the Coptic language as next source languange candidate.

CONSONANTS-VOWELS DISTRIBUTION

Latin, Greek and Coptic share another interesting feature: all the vowels are to be found within the first eight ranked glyphs. Hence by replacing them in the epigraphy with a symbolic „V“, we can observe the behaviour of all other letters. *Not all Vs are surely consonants, but all not-Vs are consonants indeed*.
After performing this operation, plenty of all-consonants triphthongs and quadriphtongs[13] are yet to be found. In Greek and Latin triphtongs are extremely rare, quadriphtpngs impossible.
This observation can be discarded only in two cases:

- All the epigraph's glyphs are consonants (for encryption reasons, or because the source language is Semitic);
- The words are anagrams;
- Some letter-groups have been set there willingly to disturb the decryption.

All these scenarios are virtually impossible, because:

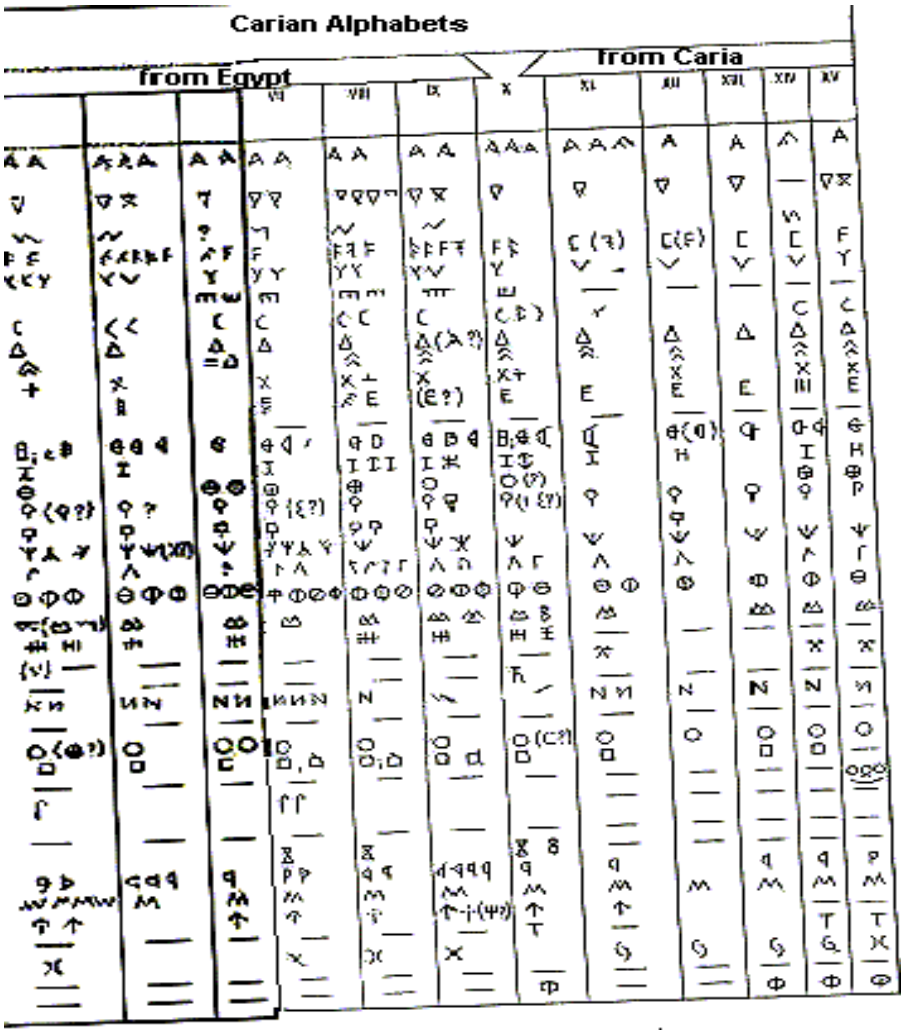
- If the original language were a Semitic one, than it would read from right to left, an option weakened by the observation that letter size *decreases* from left to right, as if the writer aimed to fit letter within the given frame;
- That the words are no anagrams is shown before through the alphabetical sorting of each word. The high TTR presented by the Coptic language is not high enough to prevent the epigraphy from displaying at least a couple of repetitions;

13 Di-, Tri-, Quadriphtongs, are respectively 2-, 3-, 4- charachters long words portions. They have no morphological meaning hence should not be confused with syllables either.

- Plenty of observations at the beginning of this paper have shown us that the epigraphy is most likely the encryption of a text written in a natural language. Nonetheless, the fact itself that the artifact is located in plain sight, entitles us to imagine that whatever information contained in the epigraphy must not be exceptionally secret, and is required to be unlocked by a group of people posessing the suitable key.

COPTIC: AN AGGLUTINATIVE LANGUAGE

The Coptic is the only one to present triphtongs and quadriphtongs, and as agglutinative language, is among the three the only one showing a very huge lexical diversity, which we have simply expressed through the type-token ratio. Moreover it is the only one showing a coherent distribution of suffixes and prefixes, as in the epigraphy, which would be impossible for Latin and very unlikely for Greek. Finally, also the word length seems to speak in favour of this language, since it is the only one presenting eighteen-, seventeen- and sixteen- letters long types.
The alphabets which more than the others resembles the epigrahy is the Carian, an Anatolian alphabet derived from Phoenician and widely present on Egyptian territory:



Picture 3: A diachronic display of Carian Alphabets

The fascination with Egyptian culture is one of the well known *leitmotiv* of the city of Naples, reifyied on the neapolitan territory in many artistic and architectonic examples.
The reason for this unusual tie has to be find in the local presence of inhabitants of Egyptian descent, namely coming from a group of alexandrine merchants establishing themselves in Naples in medieval time. On the other hand, the anthropological and geographical connection is undoubtedly underlined by a symbolic and esotheric one, a „mode“ that invested all Renaissance

Italy.
Exactly behind the mysterious epigraphy addressed in this article, in the church's cloister, the Ferrillo's tomb reports as well elements of egyptian culture:

Picture 4: Marble relief of Ferrilo tomb. First detail: the sphynx.
Second detail: the Dragon, the family's Coat of Arms

Together with the dragon, the Ferrillo family's coat of arms, on the left and right side we find the emblem of the city of Thebes, the legendary sphynx.
Finally it should be acknowledged that Coptic is a sacred language, an official language of the Church in which the large collection of gnostic, hagiographic, patristic and ancient monastic documents are edited. For the same reason, it is highly unlikely, if not impossible, that the epigraphy has been written in a common language.

RESULTS AND FINAL CONSIDERATIONS

In this paper we have thrown a close and a distant gaze at the encrypted epigraphy of Santa Maria la Nova in Naples. We have used *Bag-of-Words* approaches, such as the letter frequency calculation against big *corpora* in the main classical languages: Latin, Greek and Coptic.
Through different heuristics we have been able to theoretically strengthen a language choice, the Coptic, and weaken the other two, although Latin better provides presence of 5-lettered palindromes and better conforms to the whole context the epigraphy is placed in. On the other hand in Greek we have been able to translate a epigraphy´s line analogically, rendering it in a meaningful and plausible sentence. Moreover, the Greek's letter frequency distribution shares a remarkable similarity with the epigraph's one, which cannot either be disregarded nor discarded.
Despite all the encouraging information gathered and the clear decryption methodology drawn, a thorough decryption has not been performed yet. There are still some options to be taken in consideration, still coherent with the „1-to-1“ character substitution decryption strategy, like for instance the conjecture the epigraphy be edited in more than one language.

BIBLIOGRAPHY

Knight K., Megyesi B., Schaefer C. (2011), The Copiale Cipher, *4th Workshop on Building and Using Comparable Corpora:Comparable Corpora and the Web BUCC, 12-19,* https://dl.acm.org/doi/pdf/10.5555/2024236.2024239

Hauer B., Kondrak G. (2016), Decoding Anagrammed Texts Written in an Unknown Language and Script, *Transactions of the Association for Computational Linguistics* 4: 75–86 , https://doi.org/10.1162/tacl_a_00084

Peust, C. (2015), *On consonant frequency in Egyptian* and other languages. https://doi.org/10.11588/PROPYLAEUMDOK.00002676

Perillo M., (2018, Novembre 19) La «tomba di Dracula» a Napoli: ecco cosa si è scoperto su una misteriosa iscrizione, Il Mattino, https://www.ilmattino.it/napoli/cultura/tomba_di_dracula_napoli_scoperta_misteriosa_iscrizione-4117802.html

Köhler R. (1986), Zur linguistischen Synergetik: Struktur und Dynamik der Lexik, Quantitative Linguistics vol.31

ONLINE RESOURCES

DIZIONARIO GRECO ANTICO OLIVETTI, Retrieved June 2021 from https://www.grecoantico.com/

PYTHON- CONSTRAINT API Documentation, Retrieved June 2021 from http://labix.org/doc/constraint/

The Classical Language Toolkit (CLTK), Retrieved June 2021 from https://legacy.cltk.org/en/latest/

cltk/cltk, Retrieved June 2021 from https://github.com/cltk/cltk

Natural Language Processing with Python, Retrieved June 2021 from https://www.nltk.org/book/

Christos-c/bible-corpus, Retrieved June 2021 from https://raw.githubusercontent.com/christos-c/bible-corpus/master/bibles/Coptic-NT.xml

Digital Research in Coptic Language and Literature, Retrieved June 2021 from https://copticscriptorium.org/

Refubium, Comprehensive Coptic Lexicon: Including Loanwords from Ancient Greek v1: Version 1, Retrieved June 2021 from https://refubium.fu-berlin.de/handle/fub188/24570

Universal Declaration of Human Rights (UDHR) corpus, Retrieved June 2021 from http://research.ics.aalto.fi/cog/data/udhr/