Exploring the musicalization of texts in Gregorian Chant using Data Analytics

Hernández Benavides Miguel Angel[a] and Galpin Ixent[b]

[ab]Universidad de Bogotá Jorge Tadeo Lozano, Cra. 4 # 22-61, Bogotá D.C., COL.

ARTICLE HISTORY
Compiled September 22, 2020

ABSTRACT
Applying an algorithm for local sequence alignment in the DNA to 1,273 chants organized in pairs, a similarity percent of its texts and melodies was obtained. This generated 68,810,000 records. Using Descriptive Analytics, we conclude: in punctuation marks the sound duration is lengthened, stop words are musicalized using few pitches, semantic of text do not determine always how to musicalize it, the centonization process was influenced by the chant type and *ethos* (mode) reusing mostly phrases. These conclusions help to perform and musicalize latin texts for the Proper of the Mass. Finally, a future work in digital humanities is proposed.

KEYWORDS
Data Analytics; Gregorian Chant; Musicalization of texts; centonization; text and melody similarity.

1. Introduction

Since the second half of the 19th century, The Gregorian Chant has been in the process of restoring its melodies and interpretation. This was started in 1862 by *Dom Prosper Guéranguer* (†1875) (Combe, 2008). This work of restoration is made up of two parts which are closely related:

1. Rebuilding the true melody.
2. Find its ancient way of execution (Avedillo, 1983).

The editions of books as *Graduale Neumé* (1908), *Offertoriale Triplex* (1959), *Graduale Triplex* (1979), *Graduale Novum* (2011 and 2018), among others and the emergence of specialities such as paleography and semiology of Gregorian Chant serve as evidence that most of the restoration work has focused on rebuilding melodies exploring the musical writing that has lasted up to the present in manuscripts between 9th to 12th centuries approximately (Combe, 2008). Currently, there are several interdisciplinary projects between musicology and information technologies to contribute to the reconstruction of the melodies of Gregorian Chant. For example, projects such as (Hankinson et al., 2012) and (Helsen, Behrendt, & Bain, 2017), investigate manuscripts by processing and storing as data the information related to their object of study (Smith,

CONTACT Hernández Benavides Miguel Angel. E-mail: miguela.hernandezb@utadeo.edu.co
CONTACT Galpin Ixent. E-mail: ixent@utadeo.edu.co

2006). On the other hand, in order to "Find its ancient way of execution" it is very important to explore the text-melody relationship and the process to musicalize the texts in this repertoire. The importance of the research on this specific topic relies on the fact that the first formal declaration in the process of restoration of Gregorian Chant is:

> The rule that governs all other rules is that, pure melody apart, chant is an intelligent declamation, with the rhythm of speech, and well-phrased...(Saulnier, 2003)

The Gregorian Chant musicalizes texts to elevate them from the plane of speech to the chant in order to convey its religious meaning and sense in the Catholic worship (the liturgy). The elaboration of the main *corpus* of Gregorian Chants was completed before the end of the 8th century (Asensio, 2017) when the musical writing had not been developed yet. Therefore, much of the musicalization of texts was done by centonization (Hughes, 1987). The reuse of well-known melodies was crucial because memory and oral transmission were the only way to preserve and spread Gregorian Chants.

The study of the text-melody relationship and the musicalization of texts in Gregorian Chant has been limited by the need to do it manually because it is expensive and complicated to work simultaneously on a very large set of chants. For example, in (Le Mée, 1995) an analysis of the text-melody relationship is made only for two chants: the introit *Spiritus Domini* and the gradual *Christus factus est*. The limitation described above can be solved using Data Analytics, which is defined as:

> The process of deriving higher-level information from large sets of raw data (Calì, Wood, Martin, & Poulovassilis, 2017)

This definition suggests that the tools and techniques associated with Data Analytics enable the musicalization of texts to be processed in an automated manner. By viewing the texts and melodies of the Gregorian Chant as data, statistical information and potentially interesting patterns may be identified so as to yield valuable insights. It should be mentioned that the projects where musicology and information technologies converge can be framed within the field of research called Digital Humanities (Katz, 2005).

This paper presents the results of a first attempt to explore the musicalization of texts in Gregorian Chant using Data Analytics.

In Section 2 we present related work that provided an inspiration for this research. Musicologists and computer professionals interested in applying computer techniques to study Gregorian chant and early music have carried out projects as: a) creation of digital libraries to preserve and available to public Gregorian chant manuscripts through the internet; b) creation of very complete tables to identify and compare neumes of ancient and current writings of Gregorian chant; c) optical recognition of neumes founded in manuscripts of different traditions. Finally, we mentioned research initiatives that propose and use storing Gregorian melodies as data in order to apply Data Analytics and machine learning to find conclusions that help to solve musicological problems around of Gregorian Chant.

We decided to use Gregorian Chants of the Proper of the Mass as sample because they contain many texts musicalized with syllabic, neumatic and melismatic styles. In Section 3 we described how was obtained the sample using *webscraping* and string processing.

The scope of this research is to perform a first attempt using Data Analytics in order to find information about how the texts are musicalized in the Gregorian Chant.

Therefore, it is appropriate to carry out Descriptive Analytics to try to discover if there are patterns in the information obtained and processed.

In Section 5, we present a procedure of Descriptive Analytics for punctuation marks and Stop Words done for each pairs of chants. It consist of:

- Identify the syllables followed by a punctuation mark and describe in percentage terms the application of musical signs to lengthen their associated pitch.
- Identify the Latin Stop Words in the chants and show the number of pitches used for their musicalization.

In Section 4, we present a procedure of Descriptive Analytics using chants segmentation and similarity detection. It consist of:

- Split the text (together with its associated melody) into grammatical segments.
- Find, for each pair of chants into the sample, a percentage of similarity between texts and melodies (separately) using the *Smith-Waternam* distance algorithm (Manavski & Valle, 2008).
- Obtained the Pearson correlation coefficient between the similarity of texts and melodies.
- Show through of a scatter diagram how is the distribution of melody similarities versus the similarities of the texts.
- Show through of a scatter diagram how is the distribution of the text similarities versus the similarities of the melodies.
- Analyze in porcentual terms the reuse of musicalized texts for the segments: chants, verses, phrases, and words.

In the discussion (Section 6) we expose some points in order to think if applying Data Analytics on a wide data set of Gregorian Chants provides useful information to help the study of the musicalization of its texts and other aspects of general interest. With the results obtained in the application of Descriptive Data Analytics it is possible to think that the theology and spirituality of Gregorian Chant are also very important to -"Find its ancient way of execution".

The conclusions (Section 7) about the musicalization with one or two pitches for stop words (pronouns, articles, connectors, etc.) and the influences of the type of chant (introit, gradual, hallelujah, etc.) and *ethos* (modal classification) in the centonization, make possible to think that in Gregorian Chant it is an error to approach always the text-melody relationship from the semantics or religious message of text. Of course, this affects the performance of this music.

2. Related Work

Since the second half of the 19th Century, special interest has arisen to enrich Gregorian Chant performance by exploring its ancient written sources (manuscripts). Musical aspects of the Gregorian Chant had fallen into decline due to the rise of polyphony, music with pulse (metric), figurative writing and especially the musical and composition tendencies of subsequent epochs.

Since the second half of the 20th Century, with the emergence of computation, the Web, and information and communication technologies, research projects have been carried out to explore and digitize ancient (pneumatic) or current notations of Gregorian Chant. Its results are the sources of important information of Gregorian

Chants texts and melodies stored as data. Some of these projects are listed in Table 1.

Table 1. Research projects around exploring and digitizing ancient (pneumatic) and current notations of Gregorian Chant

Project name	**Brief description**	**Data storage format**	**Public data?**	**Related paper**
Optical Neume Recognition Project[1]	Analyze manuscript images using image processing and computational methods.	XML	No	(Helsen et al., 2017)
MEI (Music Encoding Initiative)[2]	Define guidelines for encoding music in a computer-readable structure.	XML	No	(Hankinson, Roland, & Fujinaga, 2011)
Cantus Manuscript Database[3]	Database of more than 140 manuscripts that allows searching for information and the chants contained in them.	Volpiano	No	(Lacoste & Mitchell, 2004)
Cantus Index[4]	Stored catalog of texts and melodies for the Office and Mass.	Volpiano	No	(Bezuidenhout & Brand, 2004)
Neumed and Ekphonetic Universal Manuscript Encoding Standard[5]	Data representation for the digital transcription of sources of medieval and western Byzantine chants.	XML	No	(Barton, Caldwell, & Jeavons, 2005)
SIMSSA (Single Interface for Music Score Searching and) Analysis)[6]	Way of representing music so that computers learn to recognize musical symbols in digital images of scores.	XML	No	(Fujinaga, Hankinson, & Cumming, 2014)
Search the Liber Usualis[7]	Search texts and melodies in Gregorian Chant digitized. The first results were achieved on the Liber Usualis (1961).	JSON (Apache, CouchDB)	No	(Thompson, Hankinson, & Fujinaga, 2011)
Cantus Planus[8]	Set of data for research on Gregorian chant.	Data Files *.txt	Yes	(Altstatt, 2014)
GregoBase[9]	A database of Gregorian Chant scores	GABC	Yes	(Hufflen, 2019)

Previous projects have Gregorian Chant information sources stored as data. How-

[1] http://www.cs.bham.ac.uk/~aps/research/projects/neumes/project-description.php
[2] https://music-encoding.org/
[3] http://cantus.uwaterloo.ca/
[4] http://cantusindex.org/
[5] http://www.scribeserver.com/NEUMES/main_es.htm
[6] https://simssa.ca/
[7] http://liber.simssa.ca/
[8] https://www.uni-regensburg.de/Fakultaeten/phil_Fak_I/Musikwissenschaft/cantus/
[9] https://gregobase.selapa.net/

ever, the projects Cantus Planus and GregoBase are the only ones that have downloadable public information. The Data Files *.txt of the project Cantus Planus contain a great variety of information of diverse themes (ancient antiphonaries, missals, antiphonaries, sequence texts, etc), in many cases, the files have only texts of the Chants without the melody encoded as data. The GregoBase project has files that contain the same GABC format for each chant available there. Information It is more suitable to extract a sample of chant for specifics types of repertoire (hymns, introits, antiphons, etc.) and these contain the text and melody of the chants.

Complementary to the projects above mentioned, initiatives to analyze Gregorian Chant melodies as (Van Kranenburg & Maessen, 2017) show that applying techniques of Natural Language Processing (NLP) such as *n-gram* and classification with binary trees can reach useful conclusions to the musicologist. For example, in this project, offertory melodies associated with five performance traditions from the 11th to the 13th Centuries (Gregorian, Milanese, Old Roman, Beneventan and Mozarabic) were classified and compared in order to: a) determine their origin and which membership of the offertories. They belong or could be alien to the tradition from which the melody was extracted; b) compare the five traditions with each other to quantitatively show the degree of similarity in the melodies of their offertories.

In January 2020, the not yet published project (Helsen, Daley, & Schindler, 2020) applied n-gram analysis algorithms, networks, and Recurrent Neural Networks (RNN) to quantify melodies identities for the gregorian modes looks to the nature of the gestural components of the melodies themselves.

Based on the above, it is concluded that currently in research where technological tools are used for the analysis of Gregorian chant, there is a tendency to encode and store Gregorian melodies as data in order to be able to apply computational data analytics techniques to them. In addition, this is also done because Gregorian melodies have a volume and variety of information whose manual treatment, analysis and visualization would be an almost impossible job.

3. Selection of the sample

3.1. Selected melodic styles

The Gregorian Chant musicalizes the sacred text of the liturgy (Asensio, 2017) using various melodic styles. Table 2 shows how each melodic style predominates depending on the liturgical context and the characteristics used to musicalize texts.

In the psalmodic style, the musicalization of texts is studied especially through cantillation and prosody (Chen, 1983) which consists in reusing a few melodic formulas on texts that vary. Moreover, most of the text is sung on the same musical pitch (cord). This implies that the similarity of melody between one psalm and another is very close, as the applied formulas are similar and standardized. For these reasons, this melodic style was not selected as object for this research. The syllabic, neumatic and melismatic styles were selected because they use enough melodic variations to musicalize texts. In fact, melismatic style is the most expressive resource of Gregorian Chant (Einstein, 1954).

Table 2. Melodic styles in Gregorian Chant.

Melodic style	Musicalization characteristics	Predominates in
Psalmodic	- Cantillation - Accent and rhythm of the words - Puntuation marks	- Psalms of the liturgical offices
Syllabic	- Accent and rhythm of the words - Rise and descent of the melody	- Ordinary of the Mass - Antiphons and hymns of the liturgical offices - Proper of the Mass (introit and communion)
Neumatic	- Centonization and oral transmission of melody during the period of formation of the Gregorian Chant repertoire	- Ordinary of the Mass - Proper of the Mass (introit, gradual, hallelujah, tract, offertory and communion) - Antiphons and hymns of the liturgical offices
Melismatic	- Centonization and oral transmission of melody during the period of formation of the Gregorian Chant repertoire	- Proper of the Mass (gradual, hallelujah, tract and offertory)

[a]It should be noted that it is very common to find two or more of the melodic styles listed in a single Gregorian Chant.

3.2. Selected repertoire

The chants of the Proper of the Mass (antiphon of introit, gradual, hallelujah, tract, offertory and antiphon of communion) were chosen as the sample repertoire because the musicalization of their texts involves syllabic, neumatic and melismatic styles. Consequently, they build a representative sample of the melodic styles in Gregorian Chant that were selected for this research.

3.3. Information source

There are several projects that provide information of chants stored as publical available data. For example, at the GregoBase project website[10] Gregorian Chants can be downloaded. These chants are encoded with a method known as GABC, which represents their texts and melodies using ASCII characters (Hufflen, 2019).

3.4. Sample obtained

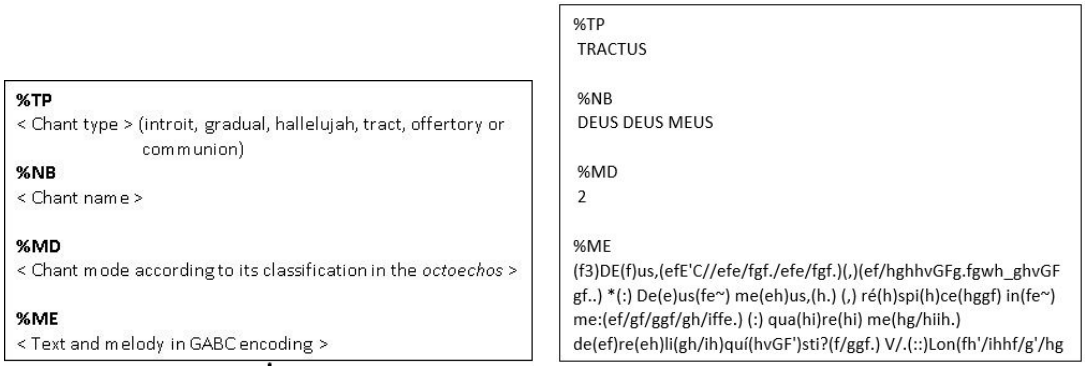
The sample of chants was obtained by implementing a *webscraping* procedure in the R language over the GregoBase website. It was done by downloading a `*.txt` file for each chant that had the option available for downloading their GABC code of the musical notation of Solesmes, in the *introitus*, *graduale*, *alleluia*, *tractus*, *offertorium* and *communio* sections of the website. The sample obtained, presented in Table 3, comprised 1,273 chants.

Table 3. Sample obtained for each type of chant.

Chant type	Frequency
Introitus	247
Graduale	190
Alleluia	305
Tractus	85
Offertorium	198
Communio	248
TOTAL	**1,273**

[10]`https://gregobase.selapa.net`

Each one of the GABC files downloaded (corresponding to the 1,273 chants) was transformed to generate another `*.txt` file, as described in Figure 1, with the following delimited sections: type (%TP), name (%NB), mode (%MD), text and melody in GABC (%ME). This was done to facilitate access the information easily in a later sequential reading of each file.



(a) Delimited sections in each file.

(b) Example for the tract *Deus Deus Meus*.

Figure 1. File **.txt* with the delimited sections generated for each chant of the sample.

3.5. Additional information for each pitch of the melodies

The additional information in Table 4 was obtained for each pitch in the chant-segment combinations (Table 6).

Table 4. Additional information stored by each pitch

Additional information	Description
It has episema	Indicates if the pitch has lengthened using a horizontal episema
It has *mora vocis*	Indicates if the pitch has lengthened using *mora vocis*
Punctuation mark	Punctuation mark in the text associated with the pitch (this value may be empty)

4. Musicalization of Punctuation Marks and Stop Words

4.1. Punctuation marks

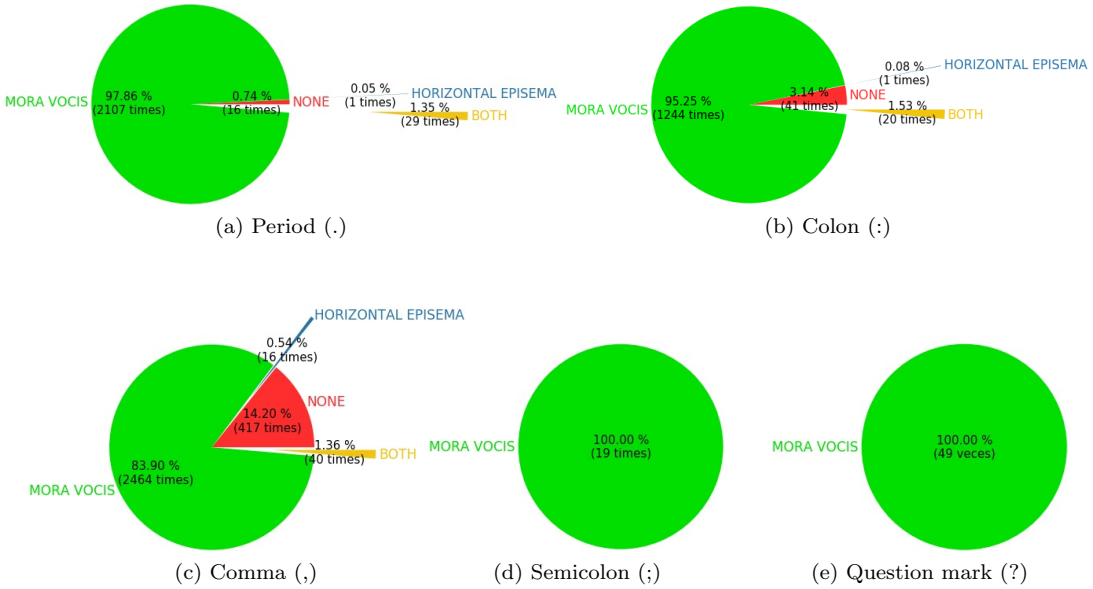
Punctuation marks in the chant texts allow: (a) phrases, sentences, etc. to be delimited to provide structure and (b) the chant to be more intelligible so as to better convey its meaning. Therefore, it seems likely that in the Gregorian Chant the musicalization of syllables taking into account punctuation marks must share some common characteristics.

6,464 syllables were found in the sample followed by one of these punctuation marks: comma, period, semicolon, colon or question mark. Table 5 and Figure 2 show the distribution (on percentage and quantity) of the use of musical symbols that lengthen the duration of the pitches associated to these syllables.

It was observed that only 3.62% of the punctuation marks do not have associated a horizontal episema or *Mora Vocis*. This indicates that lengthening the duration of the

Table 5. Use of musical symbols to lengthen the pitch in punctuation marks

Musical sign	*Mora Vocis*	Horizontal episema	*Mora Vocis* and horizontal episema	None	Total
Percentage	95.40 %	0.39 %	0.85 %	3.62 %	100.26 % ≈ 100 %
Quantity	5,883	18	89	474	6,464



(a) Period (.) (b) Colon (:)

(c) Comma (,) (d) Semicolon (;) (e) Question mark (?)

Figure 2. Detail of the use of musical symbols to lengthen the pitch for each punctuation mark.

pitch is a common characteristic (96.38% of the cases) where the syllables have some punctuation marks.

4.2. Stop words

Stop words are words that, in a certain language, do not provide by themselves important meaning to a sentence or phrase (Manning, Schütze, & Raghavan, 2008). For example, pronouns, articles and prepositions may be considered stop words. Therefore, it seems probable that if the Gregorian Chant is made based on the text to convey and enhance its message, it would not make sense to musicalize extensively (with melismatic style) the stop words, because they contribute poorly to the meaning and message of the texts.

The Classical Language Toolkit Python library[11], which has support for Latin, was used to search and group all the stop words among all the chants in the sample. As a result, 59 *stops words* were found (Table 3). The number of pitches for each stop word was illustrated through boxplots, finding that for all the stop words the the number of pitches below the third quartile (75%) is less than three; meaning that, generally the stop words are sung using one to three pitches. However, there are several atypical cases (mild and extreme) specially in chant types where the melismatic style predominates (hallelujahs, graduals, offertories and tracts), as shown in Figure 3.

Figure 4 presents the percentage of the number of notes used to musicalize stop words across the entire sample. It can be observed that, in the musicalization of stop

[11] http://cltk.org/

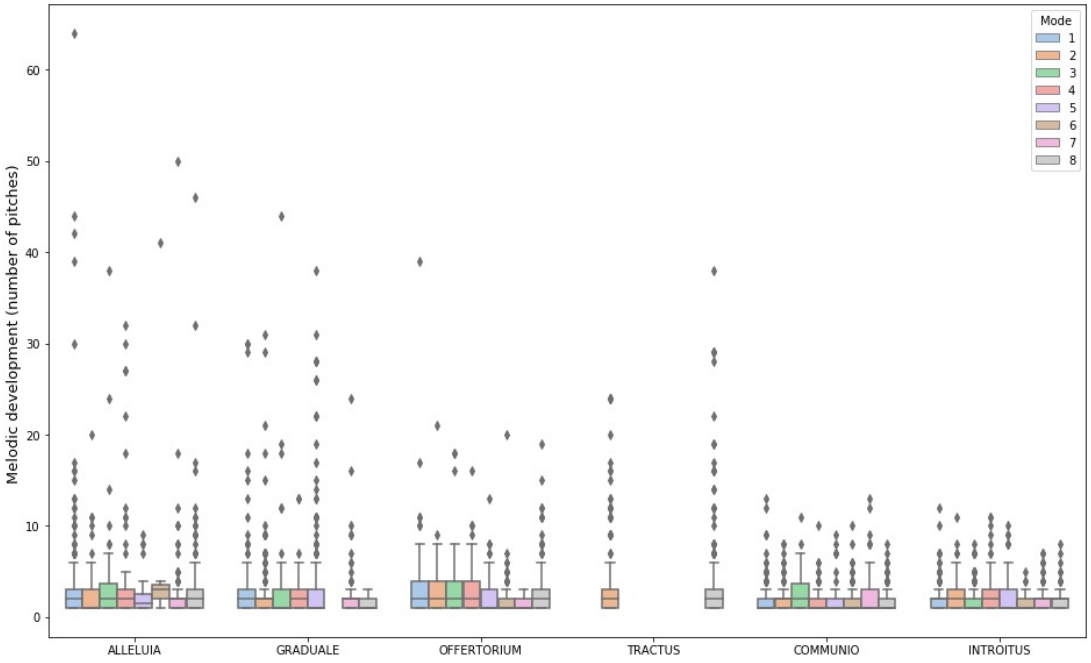


Figure 3. Number of pitches associated to stop words for each type of chant

words:

- In approximately 50% of the cases a single pitch is used, which is almost twice the number of times of two or three pitches.
- The use of two or three pitches represents approximately 10% to 25% of cases.
- The use of more than three pitches is below 5%. This means that using melisms or compound neums is atypical.

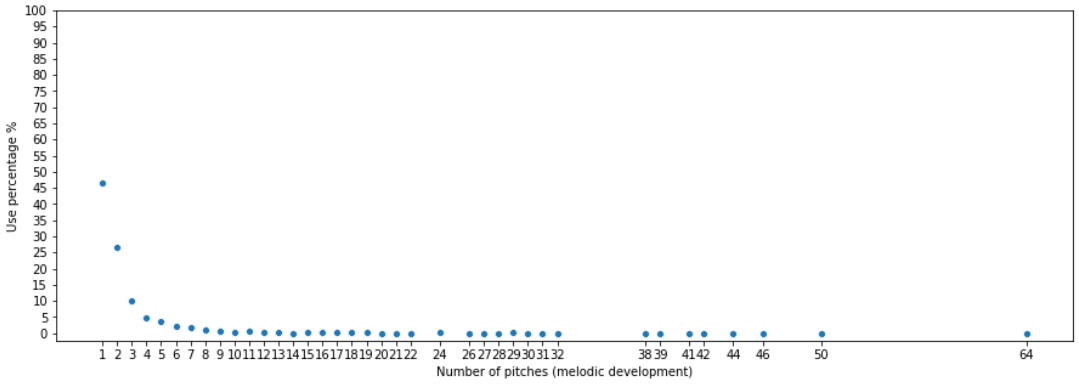


Figure 4. Percentage of the number of notes used to musicalize stop words

5. Exploring the musicalization of texts using similarity detection

5.1. Text segmentation and Smith-Waterman distance algorithm

The GABC encoding of the %ME sections (shown in Figure 1) were split using some characters as delimiters. This was done in order to segment each chant of the sample using the following units: chant (i.e., the full text of the chant), verses, phrases, words and syllables. Table 6 presents the number of records obtained according to type of unit and the characters used as delimiters.

Table 6. Segments and GABC character(s) to divide the text

Segment	GABC Character(s) to divide the text	Number of records obtained
Chant	NA	1,273
Verse	v/.	1,981
Phrase	“?” “.” “,” “:” “;”	5,969
Word	Space: “ ”	24,078
Syllable	Syllable	57,678

5.1.1. Encoding the Melody as a Character String

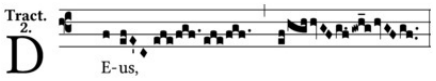
A set of alphabetical encoding (Table 7) was generated from pitch C3 to B5 using the octave index of the American Acoustic Society (ASA).

Table 7. Alphabetic encoding generated.

OCTAVE 3		OCTAVE 4		OCTAVE 5	
Note	**Alphabetic character**	**Note**	**Alphabetic character**	**Note**	**Alphabetic character**
C3	A	**C4**	I	**C5**	Q
D3	B	**D4**	J	**D5**	R
E3	C	**E4**	K	**E5**	S
F3	D	**F4**	L	**F5**	T
G3	E	**G4**	M	**G5**	U
A3	F	**A4**	N	**A5**	V
B♭3	G	**B♭4**	O	**B♭5**	W
B3	H	**B4**	P	**B5**	X

For each chant-segment combination (Table 6) the GABC melody was transformed following the values of Table 7 in order to encode the melodies as character strings. Table 8 shows the example of the melody obtained for the word *Deus* at the beginning of the tract *Deus, Deus meus.*

Table 8. Example of character string obtained for the word *Deus* (beginning of the tract *Deus, Deus meus*)

Encoding type	Example
Tetragram	
GABC	DE(f)us,(efE’C//efe/fgf./efe/fgf.) (,) (ef/hghhvGFg.fgwh_ghvGFgf..)
Character encoded (obtained from Table 7)	JIJIFIJIJKJIJIJKJIJLKLLKJKJKLKLKJKJ

5.1.2. Similarity of texts and melodies for each chant-segment of the texts

The centonization and oral transmission of the Gregorian Chant implies that there are many similar (or identical) melodic passages among various chants that have different

texts. However, the entire melody (complete sequence of pitches) in a chant identifies and distinguishes it from others; similar to how a DNA chain serves to identify every living being.

Several algorithms have been applied between pairs of DNA chains stored as data to find similar distances or regions between them (Manavski & Valle, 2008) which allows genetic relationships to be identified between species in nature. Based on this observation, the following idea is proposed: Analogous to DNA chains, distance or similarity algorithms can be applied to the Gregorian Chants to find an approximate numerical value to indicate the pairwise similarity between chant in the sample. This can be done at the level of *text* or *melody* for a given pair of chants.

5.1.3. Selecting a Distance Measure for Melody Similarity

The *Levensthein* distance algorithm calculates the minimum number of operations (minimal cost required) to transform a character string into another (Beijering, Gooskens, & Heeringa, 2008). In this case, the length of the character strings has a significant impact. It would not be effective to use this algorithm to obtain the similarity between two melodies of Gregorian Chant because there are many chants with different length and very similar melody, as can be seen in the tracts in second mode. Therefore, it was concluded that the similarity between melodies in the Gregorian Chants does not consist in determining how identical they are. Rather, it involves finding similar regions, *i.e.*, similar melodic passages.

The test displayed in Figure 5 was carried out using various algorithms used to obtain the percentage of similarity between two character strings. The complete melody of the tract *Deus, Deus meus* (one of the most extensive chants of the Proper of the Mass – 1252 pitches in total) was compared with the first 25 pitches and the 50% of the same chant. The expected result was a similarity of 100%, which would represent the total correspondence of the regions of a melody and a parts of itself.

The utility available at `https://asecuritysite.com/forensics/simstring` allows one to experiment with several algorithms to obtain the similarity between pairs of character strings. The results obtained for the tract *Deus, Deus meus* are shown in Figure 5. The *Smith-Waternam* algorithm gave the expected result. This algorithm is widely used in bioinformatics to carry out Local Sequence Alignment and to find similar regions between two DNA chains (Manavski & Valle, 2008).

For each pair of records of the chant-segment combinations (Table 6) the similarity between their texts and melodies (separately from each other) was obtained with the Smith-Waterman algorithm. The function *smith_waterman.normalized_similarity()* available in the textdistance Python library[12] was used.

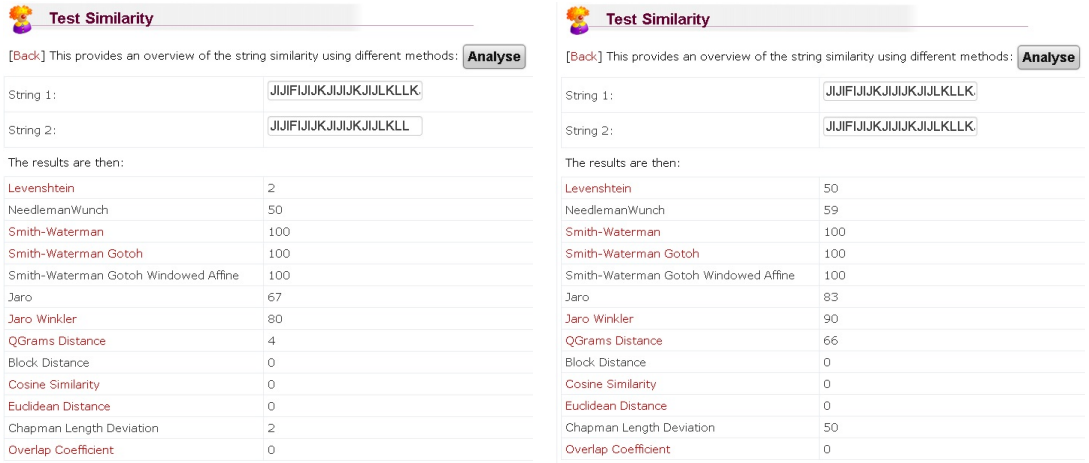
5.1.4. Exclusion of the segment: Syllable

We did not apply the Smith-Waterman algorithm on the *Syllable* segments mentioned in Table 6 due to the following reasons:

a) The difficultly of processing the high number of records that would be generated by the algorithm results, which would constitute approximately 1,663,375,842 records. This value was calculated using the formula:

$$arp = (qrs^2)/2 \tag{1}$$

[12] `https://pypi.org/project/textdistance/`



Test Similarity

[Back] This provides an overview of the string similarity using different methods: Analyse

String 1:	JIJIFIJIJKJIJIJKJIJLKLLK
String 2:	JIJIFIJIJKJIJIJKJIJLKLL

The results are then:

Levenshtein	2
NeedlemanWunch	50
Smith-Waterman	100
Smith-Waterman Gotoh	100
Smith-Waterman Gotoh Windowed Affine	100
Jaro	67
Jaro Winkler	80
QGrams Distance	4
Block Distance	0
Cosine Similarity	0
Euclidean Distance	0
Chapman Length Deviation	2
Overlap Coefficient	0

(a) Similarity to the first 25 pitches of its melody.

Test Similarity

[Back] This provides an overview of the string similarity using different methods: Analyse

String 1:	JIJIFIJIJKJIJIJKJIJLKLLK
String 2:	JIJIFIJIJKJIJIJKJIJLKLLK

The results are then:

Levenshtein	50
NeedlemanWunch	59
Smith-Waterman	100
Smith-Waterman Gotoh	100
Smith-Waterman Gotoh Windowed Affine	100
Jaro	83
Jaro Winkler	90
QGrams Distance	66
Block Distance	0
Cosine Similarity	0
Euclidean Distance	0
Chapman Length Deviation	50
Overlap Coefficient	0

(b) Similarity to 50% of its melody.

Figure 5. Similarity obtained with various distances algorithms applied to the tract *Deus, Deus meus.*

where:

arp = Total number of records to process (approximately).

qrs = Total number of records in the segment *Syllable* (Table 6).

The formula is divided by two because the Smith-Waterman algorithm has the property of being commutative.

The number of records previously obtained could not be processed with the personal computer used to perform the later phase of Descriptive Analytics. This number of records may require a cluster or *Infrastructure as a Service* in a cloud environment.

b) The syllables do not constitute meaning when they are not concatenated in the correct order to construct words. This suggests that is very unlikely that Syllables by themselves (in isolation) have a relevant impact in the musicalization of the texts in Gregorian Chant. Furthermore, the orality and analysis of this aspect have been made mainly from words (Karp, 1998), phrases and verses.

5.1.5. Limitation for the segment: Words

Using the Formula 6, the number of similarities to obtain for the words segment is 289,875,042 records. It was not possible to finish the processing for this amount of information on the computer used. Only 16% (48,227,751 records) could be processed. However, we consider that this percentage can be taken into the analysis because it is unlikely that the musicalization of texts has been done by isolated words. In the Chants of the Proper of the Mass, the isolated words do not convey a complete message.

5.1.6. Number of similarities obtained for each pairs chant-segment of the text

The total number of records in Table 9 was obtained as result of the process described in section 3.2.4. It was necessary to use *Infrastructure as a service (IaaS)* hosted by *Google Cloud*[13] and *Techila Server*[14] for the segments Phrases and Words because it would have be infeasible on a personal computer. This is mainly due to the com-

[13] https://cloud.google.com

[14] https://console.cloud.google.com/marketplace/details/techila-public/techila

putational complexity of the Smith-Waterman algorithm (Jiang, Liu, Xu, Zhang, & Sun, 2007) and the number of record pairs to be compared for each chant-segment (Table 6).

Table 9. Number of records (similarities) obtained for each segment

Segment	Records obtained
Chants	809,628
Verses	1,961,190
Phrases	17,811,496
Words	48,227,751
TOTAL	68,810,065

5.2. Similarity of texts and melodies between chants

We perform descriptive analytic to try to discover if there are patterns in the information from the similarities obtained by the segments in Table 9.

5.2.1. Correlation coefficient

The correlation coefficient for each segment (Table 6) was calculated with (a) the percentage of similarity of the texts and (b) percentage of similarity of the melodies. The results are shown in Table 10.

Table 10. Correlation between pairwise similarity of texts and pairwise similarity of melodies for each segment

Segment	Pearson Correlation
Chants	0.0255
Verses	0.0120
Phrases	0.0069
Words	0.0006

It can be seen that in all the segments the correlation coefficient is 0.0, which means that there is no relationship between the similarity of texts and melodies from a general perspective of the entire sample. However, as shown in Figure 6, when grouped by chant type and mode, an increase in the correlation coefficient is observed for all the segments where this two values are the same by pair of chants. The rise approximated of the correlation coefficient to 0.2 (+20%), 0.08 (+8%), 0.07 (+7%) and 0.01 (+1%) respectively in the chant, verses, phrases and words segments, hints that the centonization was influenced by the type and ethos (mode) of the chant.

5.2.2. Scatter diagram for similar texts

Figure 7 presents a scatter diagram showing the similarity of the melodies for each segment of Table 6, including only texts that are similar (percentage of similarity greater than or equal to 90%). It can be observed that:

- None of the graphs show any correlation between the similarities of text and melody.
- In the region where the similarity of melodies is the largest, the type of chant and mode are usually the same. This indicates that similar (or equal) texts were centonized if they belong to the same type of chant (introit, gradual, hallelujah, etc.).

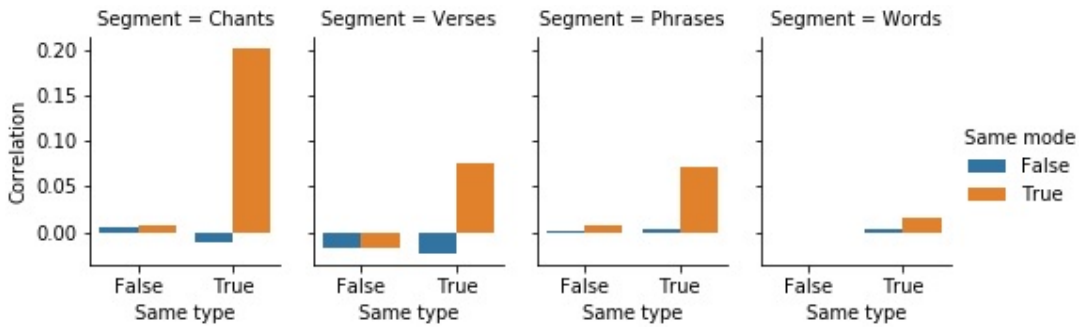


Figure 6. Text and melody similarity correlation adding the type and mode of the chants

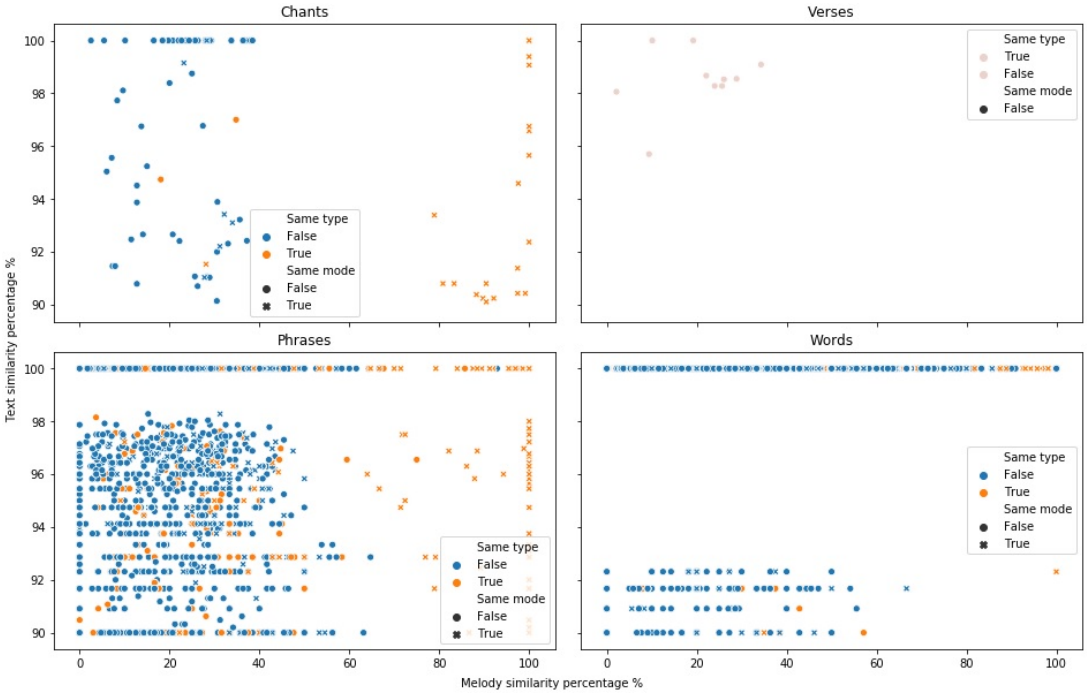


Figure 7. Melody similarity distribution for texts whose similarity is greater than or equal to 90%

5.2.3. Scatter diagram for similar melodies

A scatter diagram (Figure 8) of the similarity of texts was elaborated. For each segment of the Table 6, choosing only the melodies that are similar (percentage of similarity greater than or equal to 70%). It can be observed that:

- None of the graphs show any correlation between the text and melody similarity.
- In the graphics for the Chants and Verses segments, the same mode predominates and the types of chants are equals although the percentage of similarity among texts varies from 0% to 100%. This indicates that many different texts were simply musicalized by applying centonization based on the type chant and mode. Therefore, in many Chants of the Gregorian repertoire the text-melody relationship is not always based in the semantics or religious message of the text.
- In this region, the graph of the Phrases segment is much more populated. This allows us to formulate the hypothesis that the centonization was mostly done by

phrases of the text.

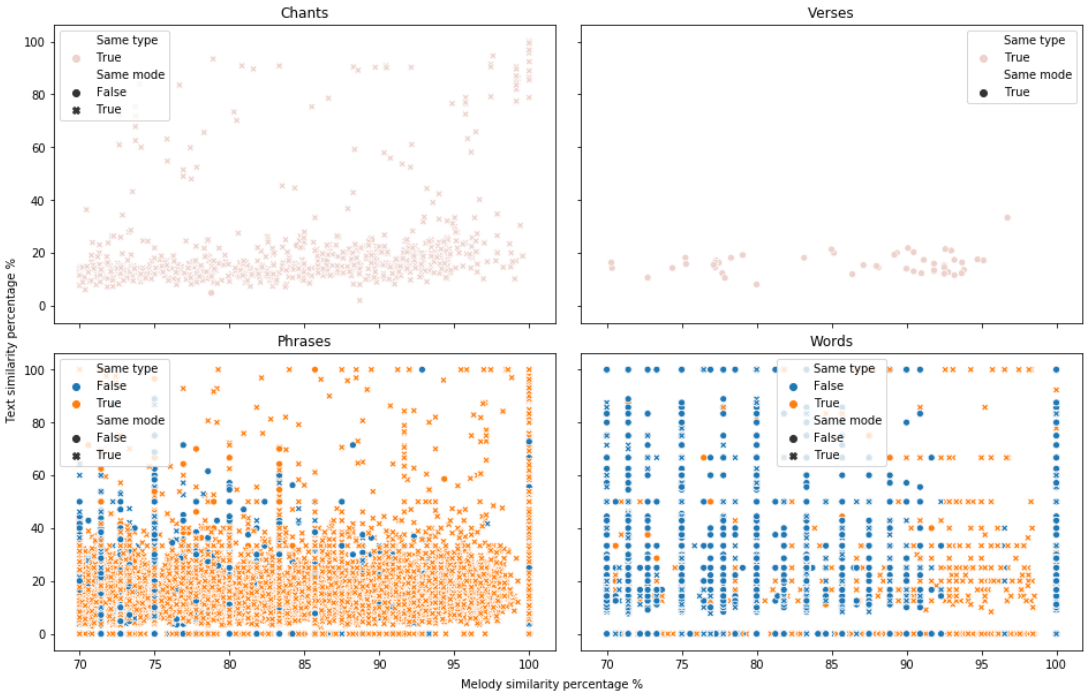


Figure 8. Similarity distribution greater than or equal to 70% between melodies.

5.2.4. Reuse of already musicalized texts

For each segment in the Table 6, the records with text similarity greater than or equal to 90% and the same type and mode were selected. Table 11 shows the percentage of cases located in each range by deciles of the melody similarity.

Table 11. Percentage of cases for each range of melody similarity (centonization)

Similarity range of the melodies	**Chants**	**Verses**	**Phrases**	**Words**
0-10 %	0.00 %	N/A	1.36 %	0.81 %
11-20 %	0.00 %	N/A	4.88 %	2.52 %
21-30 %	3.12 %	N/A	5.07 %	9.01 %
31-40 %	0.00 %	N/A	4.10 %	20.05 %
41-50 %	0.00 %	N/A	2.53 %	10.86 %
51-60 %	0.00 %	N/A	0.58 %	1.37 %
61-70 %	0.00 %	N/A	1.17 %	0.49 %
71-80 %	6.25 %	N/A	1.95 %	0.35 %
81-90 %	18.75 %	N/A	2.14 %	0.07 %
91-100 %	71.87 %	N/A	73.82 %	13.53 %

[a]The Verses segment does not present results because it does not have texts with similarity greater than or equal to 90%.

[b]The sum of percentages in the Words segment does not reaches 100% due to the limitation described in Subsection 5.1.5

6. Discussion

Our results shown that in the Gregorian Chant the text-melody relationship is not always based on the semantics or meaning of the text. We can postulate this because:

- There are atypical cases where the Stops Words were musicalized with extensive melisms. Some examples are:
  - Hallelujah - Senex puerum (64 pitches in the Stop Word *autem*)
  - Hallelujah - Domine in virtute puerum (41 pitches in the Stop Word *et*)
  - Tract - Qui regis Israel (38 pitches in the Stop Word *nos*)
- There are quite a few cases where the similarity of the text is minimal, but the melody similarity is maximum. Some examples are:
  - The tracts - *Stabat Sancta Maria* and *Vere languores* have similarity in the melody of 100% and 0% in the texts: *et videte* and *super eum* respectively.
  - The hallelujahs - *Accedite* and *Inebriabuntur* have similarity in the melody of 93.2% and 11.59% in the text (its verses): *Accedite ad eum, et illuminamini, et facies vestrae non confundentur* and *Inebriabuntur ab ubertate domus tuae: el torrente voluptatis tuae potabis eos* respectively.
  - The Graduals - *Justus ut palma* and *Requiem aeternam* have similarity in the melody of 92.71% and 10.89% in the text: *Justus ut palma florebit: sicut cedrus libani multiplicabitur in domo domini. Ad annuntiandum mane misericordiam tuam, et veritatem tuam per noctem* and *Requiem aeternam dona eis domine: et lux perpetua luceat eis. in memoria aeterna erit justus: ab auditione mala non timebit* respectively.

The fact that in approximately 75% of the cases the stop words are musicalized with one or two pitches, allows us to launch the hypothesis that, to musicalize texts in Gregorian chant, it is valid to use a reduced number of pitches in words that do not have much importance within the message to be transmitted. On the other hand, the most important words or phrases could have long melisms (in the melismatic style) or several notes (in the neumatic style), especially when they transmit main ideas according to the semantics of the text or religious message. For example:

- The word *Alleluia* on paschal time.
- The hallelujah - *Deus, qui sedes* has 82 pitches in the word *Thronum.*
- The hallelujah - *Pascha nostrum* has 61 in the word *immolatus.*

Centonization can be applied to a different text as long as it is used in the same type of chant (introit, gradual, hallelujah, etc.) and mode. This suggests that it is possible to think that there is a strong relationship between melodies in a certain mode (*ethos*) and the liturgical context regardless of the text musicalized. For example, among the tracts in Lent, the octave of Easter, etc. Furthermore, the above suggests that although the Gregorian Chant has a liturgical context and sense, the musical coherence (the character or *ethos*) will not lose importance to the centonization of texts. However, it should be clarified that the more important criteria was the liturgical moment. This can be evidenced as in the graph "Chants" in Figure 8 all chants have the same type and only some cases are presented with different mode.

The use of centonization to musicalize texts, was made by keeping in mind the part of the liturgy (introit, hallelujah, tract, etc.) for which the new text will be used. For example, two equal (or similar) texts should not be centonized (have the same melody) if the text to be musicalized will be used in another part of the liturgy.

It might be thought that most of the centonization used to musicalize similar texts

in Gregorian chant was made by phrases. For example, in Table 11, 73.82 % was the highest percentage of cases found for a similarity of melody greater than 90 % and corresponds to the segment called *Phrases*. Some examples are:

- The graduals - *Locus iste Domine* and *Protector noster* have in common the musicalization of the phrase *exaudi preces servorum tuorum.*
- The tracts - *Ego diligentes* and *Meum est consilium* have in common the musicalization of the phrases *et qui mane vigilant me* and *propitius esto peccatis nostris.*
- The offertories - *Justitia indutus* and *Eo quod* have in common the musicalization of the phrases *oculus fui cæco, et pes claudo* and *pater eram pauperum.*

Using Data Analytics to explore the musicalization of texts in Gregorian Chant may help to "Find the ancient performance of Gregorian Chant". However, theology and spirituality of Gregorian Chant are important in the reconstruction of their original performance because the Descriptive Analytics applied reveals that the process of musicalization of the text is not completely based on text semantics, musical aesthetics or on a merely methodical process. And as expressed in (Asensio, 2017):

> In a culture of the concrete such as the medieval, the sounds cannot be captured from the air and written. What could be represented is something different than the sound: this is its materiality, not its essence.

Then, in the search for that essence, the interdisciplinarity of digital humanities is very useful because the technique and logic of analytical process on a wide set of information, or the musicological methods, are not by themselves sufficient (Hourlier, 1995). It is also necessary to understand the spirituality of the texts that are musicalized in the Gregorian Chant from the theology or philosophical-aesthetic-poetic aspects (Villegas, 1999).

From a point of view not purely musicological, but practical, the conclusions obtained, also may help to:

- The work of performers, composers and musicologists into their respective areas of specialization or research in Gregorian Chant. (Toloza, 1961).
- Present tools to elaborate or apply centonization to the new latin texts of the Proper of the Mass (taken from *editio typica* of roman missal) that have not been musicalized as Gregorian Chant. (Smith, 2006).
- Contribute to the research and general interest in the study of the language and music relation: which as stated in (Dydo, 1983):

  > The language and music relationship have been for us directly an inheritance of Gregorian Chant.

7. Conclusions

From the Descriptive Analytics applied, it is concluded that for the musicalization of texts in Gregorian Chant:

- The pauses of the spoken voice in the reading of words with punctuation marks, are musicalized in the Gregorian Chant lengthening the duration of the pitch in the syllable that precedes the sign. It was found that this occurs in 96.38% of cases, which suggests that it is very important to keep punctuation marks in mind

when performing Gregorian melodies in syllabic, neumatic and melismatic styles. This makes sense, because the punctuation marks give structure and facilitate the transmission of the message text.

- There are quite a few cases where the stop word were musicalized with extensive melisms.
- Stop word are mostly musicalized with one or two pitches in approximately 75% of the cases, which may be considered as an attempt to downplay them by reducing as much as possible the number of pitches used on its performance.
- The Pearson correlation coefficient between the similarity of texts and melodies increases in approximately 20% when the chants have the same type and mode.
- When the melody similarity reaches 91% to 100%, the highest percentage of similar texts cases (73.82 %) is present in the Phrases segment. This suggests that the phrases were the most predominant segment where the previous existing musicalization (melody) of the texts were reused (centonization). This is the case only when the chants have the same type and mode.

8. Future work

Another exploration of the musicalization of texts in the Gregorian Chant could be made from a semantic approach. To achieve this, some musical passages would be chosen already analyzed by some authors (especially musicologists or religious) where they expose the relation between the melody and the semantics or religious message of its text. This could constitute an ensemble of information to train a classifier model that is able to ascertain the level of text-melody relationship in a specific fragment of this music. It would use *machine learning* and *Natural Language Processing* (NLP).

As the Gregorian chant was mostly musicalized using centonization, one may think that one melody may be older than another if its relationship with the semantics or religious message of the musicalized text is greater. To formalize a hypothesis such as the one stated above, the collaboration of musicologists and experts in Gregorian Chant is needed to suggest the information that must be added to the sample of chants to apply Data Analytics, *e.g.*, historical information to avoid the absence of a time frame in analytics on this topic (Karp, 1998). These experts will also validate the coherence of the results obtained with the conclusions of previous musicological research on this topic.

This future work may be considered as a research project in the field of digital humanities (Berry, 2012).

References

Altstatt, A. (2014). Reviews: Cantus planus regensburg, corpus antiphonalium officii-ecclesiae centralis europae, cantus: A database for latin ecclesiastical chant, global chant database, and the cantus index. *Journal of the American Musicological Society*, *67*(1), 267–285.

Asensio, J. C. (2017). *El canto gregoriano: historia, liturgia, formas.* Alianza Editorial.

Avedillo, M. (1983). Fabriciano: Canto gregoriano. estudio teórico y práctico. *Zamora: Fama [1983].*

Barton, L. W., Caldwell, J. A., & Jeavons, P. G. (2005). E-library of medieval chant manuscript transcriptions. In *Proceedings of the 5th acm/ieee-cs joint conference on digital libraries* (pp. 320–329).

Beijering, K., Gooskens, C., & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm. *Linguistics in the Netherlands*, *25*(1), 13–24.

Berry, D. M. (2012). Introduction: Understanding the digital humanities. In *Understanding digital humanities* (pp. 1–20). Springer.

Bezuidenhout, M., & Brand, M. (2004). The "cantus index gui": A visual interface for the creation of electronic plainchant resources. *Studia Musicologica Academiae Scientiarum Hungaricae*, *45*(1-2), 3–16.

Calì, A., Wood, P., Martin, N., & Poulovassilis, A. (2017). *Data analytics: 31st british international conference on databases, bicod 2017, london, uk, july 10–12, 2017, proceedings* (Vol. 10365). Springer.

Chen, M. Y. (1983). Toward a grammar of singing: tune-text association in gregorian chant. *Music Perception: An Interdisciplinary Journal*, *1*(1), 84–122.

Combe, D. P. (2008). *The restoration of gregorian chant: Solesmes and the vatican edition.* CUA Press.

Dydo, S. (1983). Surface relations between music and language as compositional aids. *Journal of New Music Research*, *12*(4), 541–556.

Einstein, A. (1954). The conflict of word and tone. *The Musical Quarterly*, *40*(3), 329–349.

Fujinaga, I., Hankinson, A., & Cumming, J. E. (2014). Introduction to simssa (single interface for music score searching and analysis). In *Proceedings of the 1st international workshop on digital libraries for musicology* (pp. 1–3).

Hankinson, A., Burgoyne, J. A., Vigliensoni, G., Porter, A., Thompson, J., Liu, W., ... Fujinaga, I. (2012). Digital document image retrieval using optical music recognition. In *Ismir* (pp. 577–582).

Hankinson, A., Roland, P., & Fujinaga, I. (2011). The music encoding initiative as a document-encoding framework. In *Ismir* (pp. 293–298).

Helsen, K., Behrendt, I., & Bain, J. (2017). A morphology of medieval notations in the optical neume recognition project. *Arti musices: hrvatski muzikološki zbornik*, *48*(2), 241–266.

Helsen, K., Daley, M., & Schindler, J. (2020). *Quantifiying the melodic identifies of medieval modes.*

Hourlier, D. J. (1995). *Reflections on the spirituality of gregorian chant.* Paraclete Press.

Hufflen, J.-M. (2019). Antichi sistemi di notazione musicale. *ArsTEXnica*, 42.

Hughes, D. G. (1987). Evidence for the traditional view of the transmission of gregorian chant. *Journal of the American Musicological Society*, *40*(3), 377–404.

Jiang, X., Liu, X., Xu, L., Zhang, P., & Sun, N. (2007). A reconfigurable accelerator for smith–waterman algorithm. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *54*(12), 1077–1081.

Karp, T. (1998). *Aspects of orality and formularity in gregorian chant.* Northwestern University Press.

Katz, S. N. (2005). Why technology matters: the humanities in the twenty-first century. *Interdisciplinary science reviews*, *30*(2), 105–118.

Lacoste, D., & Mitchell, A. (2004). The cantus database: Progress report. *Studia Musicologica Academiae Scientiarum Hungaricae*, *45*(1-2), 119–130.

Le Mée, K. (1995). *El canto gregoriano: su historia y sus misterios.* Temas de hoy.

Manavski, S. A., & Valle, G. (2008). Cuda compatible gpu cards as efficient hardware accelerators for smith-waterman sequence alignment. *BMC bioinformatics*, *9*(S2), S10.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval.* Cambridge university press.

Saulnier, D. D. (2003). Gregorian chant. *A Guide to the History and Liturgy*.

Smith, R. A. (2006). Roots into the future: Recovering gregorian chant to renew the church's voice. *Theology Today*, *63*(1), 48–54.

Thompson, J., Hankinson, A., & Fujinaga, I. (2011). Searching the liber usualis: Using couchdb and elastic search to query graphical music documents. In *Proceedings of the 12th international society for music information retrieval conference.*

Toloza, D. L. (1961). Prolemática de la actual investigación gregoriana (ii). *Revista Musical Chilena*.

Van Kranenburg, P., & Maessen, G. (2017). Comparing offertory melodies of five medieval christian chant traditions. In *Ismir* (pp. 204–210).

Villegas, L. P. (1999). La estética musical en el medievo cristiano o un viaje estético a las formas musicales gregorianas. *Revista Española de Filosofía Medieval*, *6*, 77–103.