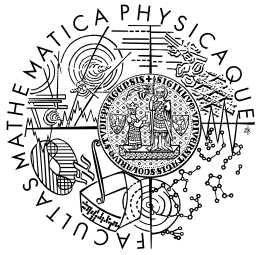

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Name Surname

Thesis title

Name of the department

Supervisor of the bachelor thesis: Supervisor's Name
Study programme: study programme
Study branch: study branch

Prague YEAR

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication.

Title: Thesis title

Author: Name Surname

Department: Name of the department

Supervisor: Supervisor's Name, department

Abstract: Abstract.

Keywords: key words

Contents

Introduction

Imagine, for a moment, that you are transported in time to 10th-century Europe. It is Sunday, and the weekly mass is just starting. Look around yourself. The tall ceilings of the impressive church were built so to bring you closer to God. The stained glass windows let through just enough light so that it is not completely dark. Can you smell the incense? Everything in the environment around you reminds you that you are taking part in a sacred ritual. Listen to the priest; when he recites the prayers, his speech does not flow in a natural way. Instead, the entire text is intoned on a single note, except for the slightly inflected ends of clauses. Sometimes, the choir replaces the priest, singing more elaborate monophonic melodies. The words they are singing are Latin, but even if you do not understand them, you know what their purpose is: to celebrate the deity. Their voices echo in the stone church, creating an otherwordly experience.

What you are hearing is called Gregorian chant. It is one of the earliest forms of music preserved in written form, and the largest preserved body of medieval music. The earliest preserved fragments of written notes date back to the 9th century, although texts from as early as the eighth century have been found.. Its name, Gregorian, references Pope Gregory the Great, however, his relation to the chant is not entirely clear.

Gregorian chant is not the only type of chant. In the early centuries after Christianity spread across Western and Eastern Roman Empire, new forms of worship started being developed. The most obvious differences were between the West and the East, which had multiple cultural centers such as Constantinople, Jerusalem, or Alexandria. However, liturgies varied in the West as well, from Rome to Milan to the Iberian peninsula to Gaul. Each center developed their own tradition, including their own type of chant.

(TODO: why is the Roman tradition everywhere now?)

Gregorian chant is an integral part of the Roman church, and has been so for centuries. It is the monophonic music (i.e. single-voice) sung during liturgies. Here, liturgy means chant sung during Christian worship. Unlike in the Eastern churches, where the term is reserved for the Eucharist, liturgy includes both the Mass and the Divine Office in the Roman church.

Mass is the service most familiar to most believers. It can be divided into several parts, all leading up to the most important one, the act of communion. This act commemorates the Last Supper, Jesus's last meal with his disciples before his execution. During the communion, bread, representing Christ's body, and wine, representing his blood, are given out. During the course of the Mass, multiple different chants are sung. *Introitus*, meaning 'entrance', is sung at the beginning of the service while the priest and his assistants are walking to the altar. *Tractus* is a chant that is sung during Lent, i.e. the period between Ash Wednesday and the Saturday before Easter. Outside of this period, *alleluia* replaces *tractus* and is followed by *sequentia* on the most important feast days.

The other part of the liturgy, besides the Mass, is the Divine Office, also called Liturgy of the Hours or canonical hours. It is the set of chants sung during services at different times of the day, for example *Vespers* in the evening or *Lauds* in the morning. The office consists largely of singing psalms, of which there are 150,

all sung on different days and hours. Psalms were usually preceded and followed by antiphons, which differ not only by the day of the week, but also depending on the place. Each day of the week had an allocated set of antiphons, hymns and responsories, while responsories were assigned to the different Sundays. Additionally, important feasts had their own set of chants to be sung.

The liturgical year contains several feasts. Some feasts are fixed on a specific date. Feasts associated with a specific saint are an example of those. For example, John the Baptist is celebrated on the 24th of June, Michael on the 29th of September, and All Saints on the 1st of November. Other fixed-date feasts include Christmas Day (25 December), Epiphany (6 January), and others. Such feasts can fall on any day of the week, and if they fall on a Sunday, they will take precedence over it. On the other hand, there are also feasts that are fixed to a specific day of the week, the most important of them being Easter Sunday, the day when Christ rose from the dead. Each feast has a different set of chants that can be completely original.

The individual chants differ in several criteria, the first one being the mode. Mode is the system of pitch organization, somewhat similar to modern-day scales. Melodies are classified into one of eight modes according to their last note, called *finalis*, and their range. Most chants end on one of the notes *D*, *E*, *F*, or *G*. These four notes determine four pairs of modes. The melodies were further classified depending on whether they moved mostly in the range above the *finalis*, in which case it would be classified as the *authentic* mode of the pair, or in the range around the *finalis*, which means it is classified as the *plagal* mode. Some types of chant tend to occur mostly in a specific mode.

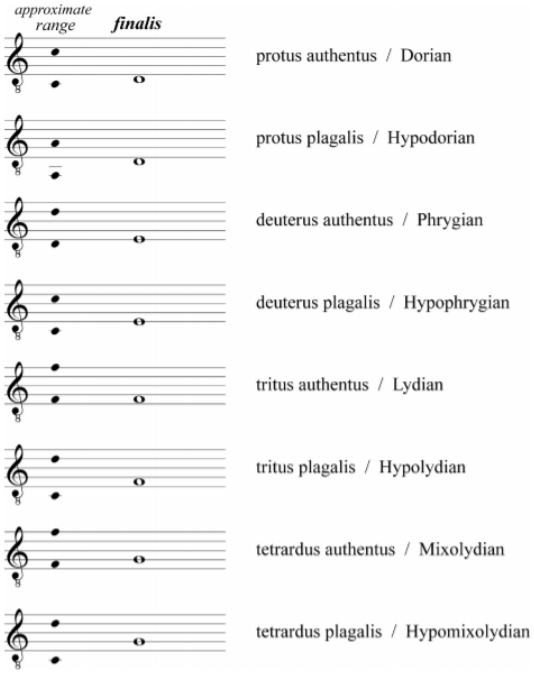

Figure 1: List of modes, their ranges and finalis. [Hiley, 2009, p. 44]

Another criterion is the complexity of chants, that is, how elaborate the melody is. On the one side of the spectrum, there is a one-to-one syllable to

note correspondence. Antiphons and some hymns are genres close to this text-setting. The other extreme is melismatic style. Melisma is a long vocalization of a single syllable, therefore melismatic melodies are more ornate. Some of the more melismatic genres are the gradual, tract and offertory.

We have already mentioned several different genres of chant. Antiphons are chants sung to frame a psalm during the Office hours. They are relatively short and simple, somtimes consisting of only two phrases. Responsories are sung during the Night Office. Each has a main section and a verse. They are melodically very rich and are on of the most impressive forms of chant. The number of chants sung during the Mass is lower, as there will usually only be one introit, one gradual, etc. during a service.

It is important to note that text and melody do not form unique pairs. Instead, each text can be sung to multiple different melodies and multiple texts can be sung to one melody.

It is clear that the annual cycle is very complex and the amount of chants is abundant. Therefore, it is not surprising that while the chants during the services themselves were sung from memory, they were written down in books. Each church and each convent had their own manuscripts, which is the reason of their abundance.

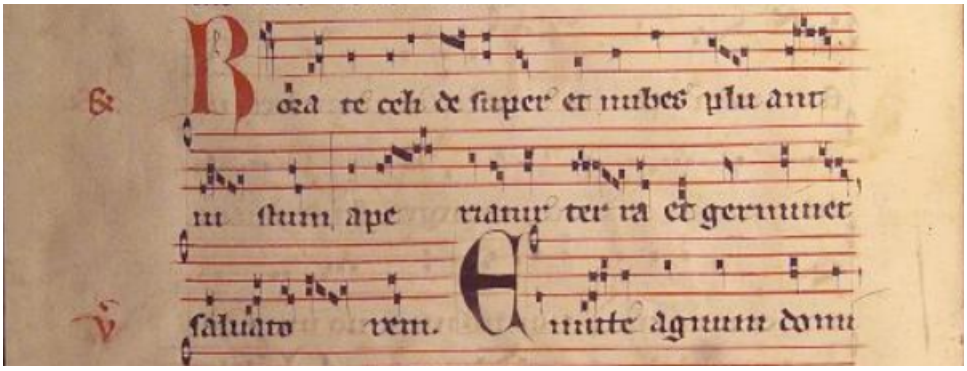

Figure 2: Example of a chant in a manuscript. [Lacoste et al., id 007553]

1. Related work

With the advent of computers, the field of Music Information Retrieval (MIR) emerged. Research in this interdisciplinary field focuses on extracting information from musical notation using computer science methods, such as signal processing or machine learning. Its applications vary widely, from recommender systems, to automatic audio transcription, to music generation. MIR encompasses all different kinds of music, regardless of their location, age, or function. Researchers have developed a multitude of software tools that facilitate music analysis, irrespective of what type of music it is. One of such toolkits is *music21* [Cuthbert and Ariza, 2010], a Python package able to encode musical notation as Python objects and perform analysis on large datasets.

The study of plainchant using computational methods has not been done extensively. The main research tool for musicologists in this field is the Cantus Index [Lacoste and Koláček]. It is an online index of chants from several different chant databases, providing researchers with a common API for all of them. The entries in Cantus Index only contain four data fields: full-text, genre, feast (not required), and Cantus ID, which is automatically assigned to newly added chants. The tool is also able to search for melodies in the original source and even provides search-by-melody functionality. There are ten databases indexed in the catalogue, the largest of which is the Cantus database [Lacoste et al.].

Cornelissen et al. [2020a] developed *chant21*, a Python package able to convert two standard melodic notations, *volpiano* and *gabc* to a *music21* object, therefore making it easier to study Gregorian chant computationally. The data they used were scraped from Cantus database [Lacoste et al.] and GregoBase [Berten and contributors] and released as CantusCorpus and GregoBaseCorpus, respectively. Finally, they performed two case studies using the package. In the first one, they confirmed the melodic arch hypothesis [Huron, 1996], which had previously only been studied manually. Second, they analyzed the relation between differentiæ and antiphon openings [Shaw, 2018] and found that it differs accross modes.

Some of the computational research into plainchant has been centered on mode classification. Huron and Veltman [2006] used pitch class profiles to classify modes. They created a pitch-class distribution for each of the eight modes, and used these classes to classify previously unseen data. Cornelissen et al. [2020b] compared three approaches to mode classification: classical approach, which classifies chants based on the final pitch, range, and the initial pitch; profile approach, which was largely inspired by Huron and Veltman [2006]; and distributional approach, which focuses on the melodic aspect of mode. The authors chose various segmentations and representations of chants and used a tf-idf vector model to classify mode. The study found that we can accurately classify mode even when we discard all absolute pitch information, the melody contour contains enough information on its own.

A considerable amount of research has been done into the evaluation of melodic similarity, albeit not for Gregorian chant specifically. Wickland [2017] provides an overview of the methods. He mentions edit distance, Markov chains, and geometric measurements as the most widely used ones. Park et al. [2019] used an adapted edit distance metric to calculate the similarity of two melodic sequences

by first calculating the similarity for all segments of each of the sequences and then scaling them by a weight function depending on the segment length, which yielded them what they call a multi-scale similarity stack. The overall similarity was obtained by averaging its values. Then they used the MSS stack to create a visualization that takes on the shape of a trapezoid that shows which segments of two sequences are the most similar.

Bountouridis et al. [2017] argue that methods originally developed for bioinformatics have a great potential to be applied to music. They offer analogies for bioinformatics concepts found in musicology. For example, they liken DNA and proteins to melodic sequences, homologues (proteins that have the same ancestor) to song covers, evolution to oral transmission, etc. They claim that despite the similarities, MRI has not leveraged the full potential of bioinformatics methods. In their article, they focus on modelling melodic similarity using multiple-sequence alignment (MSA) algorithms, therefore not relying on heuristics, as opposed to previous works. Their results revealed that the MAFFT algorithm yields the best alignemnt, which can be attributed to the algorithm using gap-free segments as anchor points, therefore partitioning melodies into more meaningful segments than other algorithms.

The general algorithm for calculating pairwise sequence alignment is the Needleman-Wunsch algorithm [Needleman and Wunsch, 1970]. It uses dynamic programming to break down the problem into smaller problems. Given two sequences, it starts aligning them from the beginning. At each point, the algorithm checks whether the two sequences match in the current position, and if not, whether it will leave the elements mismatched or insert a space. In essence, all possible alignments are computed and scored and the best one is chosen. The algorithm always yields an optimal alignment, therefore it is used when the quality of the alignment is important. However, because of its time complexity, it is unsuitable for many applications.

Unlike pairwise sequence alignment, multiple sequence alignment has been shown to be NP-complete [Wang and Jiang, 2009]. As such, there is no practical way of computing an optimal MSA and we must instead rely on heuristics to obtain a sufficiently good alignment.

Notredame [2007] provides an overview of modern multiple sequence alignment algorithms. According to him, the most frequently used algorithms use the progressive approach, where a guide tree is estimated from unaligned sequences and then pairwise alignment algorithms are used to find the MSA following the tree. He notes that the scoring methods of the pairwise algorithm are essential. There are two main groups of scoring methods: matrix-based algorithms, where a substitution matrix is used to determine the cost of replacing one symbol with another, and the consistency-based methods, which use a collection of global and local alignments to calculate a position-specific substitution matrix. The author claims that the best methods yield indistinguishable results, except for remote homologs with less than 25% identity.

T-Coffee [Notredame et al., 2000] uses the progressive approach described above. It was the first algorithm that used a preprocessed collection of alignments to create a library that helps create the guide tree. The library is generated using both global and local pairwise alignments. Thanks to this approach, T-Coffee minimizes the errors made in the first stages of building the MSA, which is a

shortcoming of many previous algorithms, as these errors tend to persist. They combined precomputed local and global alignments and create a function that assigns a weight to each pairwise alignment depending on how consistent the pair of residues is with the residue pairs from all other alignments. This process leads to a significant improvement of the results.

MAFFT [Katoh et al., 2002] further improves on other methods by using Fast Fourier transform to identify homologues fast. In addition, the authors propose a simplified scoring system that reduces CPU time while maintaining its accuracy. The authors' results showed a performance 100 times better than that of T-Coffee.

2. Data

Our main source of data is the Cantus database (Lacoste et al.), one of the databases indexed in the Cantus Index. The database serves as a digital archive of chants, each entry containing information about its source, liturgical occasion, mode, and others. Work on the project started in the late 1980s, and to date, around 500,000 individual chants from approximately 150 manuscripts have been indexed. Each entry is transcribed manually and undergoes a thorough examination before publishing (Lacoste [2012]).

We are using a scraped version of the Cantus database released as CantusCorpus (**?**chant21)). Unlike the Cantus database which is continuously being updated and is therefore unsuitable for computational study, the corpus is versioned, therefore each version always contains the same data. We are using version 0.2 released in July 2020 which contains 497071 entries. The corpus is available for download in CSV format.

2.1 CSV

CSV is one of the most common formats for tabular data. The abbreviation stands for *comma-separated values*. As the name suggests, the format uses commas to separate columns (although other separators, such as a semicolon, can be used as well to allow for simpler parsing in case that the data frequently contains commas that would otherwise need to be escaped), while the individual rows are separated by a line break. The data is stored as plaintext, which makes it easily readable. Parsing CSV files becomes more complicated when the data contains column and row separators inside fields; in that case quotation marks or esape sign has to be used. There exist many well-designed parsers, one such parser is the Python module simply called *csv*. This application uses the module *pandas* to parse CSV files, which in turn uses the *csv* module.

2.2 Database fields

The following table represents the data fields in the database.

Data field	Description
id	automatically generated id in the database
corpus_id	human-readable id identifying the chant in the CantusCorpus
incipit	incipit (the first few words) of chant
cantus_id	id identifying the chant in the Cantus Index
mode	mode of the chant
finalis	the final note of the chant
differentia	FILL IN
siglum	FILL IN
position	liturgical role of the chant
folio	page of the manuscript where the chant is found
sequence	order in which the chant is found in the folio

marginalia	clarification about the location of the chant
cao_concordances	FILL IN
feast_id	id of feast
genre_id	id of genre
office_id	id of office
source_id	id of source
melody_id	id of melody by which it can be found in the Cantus Index
drupal_path	URL of the chant on the Cantus database website
full_text	full text in a standardized spelling
full_text_manuscript	full text in the manuscript spelling
volpiano	transcription of the melody
notes	indexing notes

Table 2.1: List of database fields

2.3 User-defined data

The application enables user to upload their own dataset. In doing so, the user should upload a CSV file with the fields described above, with some clarifications:

- *corpus_id* is a name invented for this application. The original CSV file has this column listed as *id* and to maintain consistency, so should every user-defined file.
- The column *id* in this application is generated on data upload. This column should not be present in the uploaded data.
- There can optionally be an unnamed column in the first position that will be dropped. It is so in case the user generates data from a database and doesn't remove its original ids.
- The fields *full_text* and *volpiano* should contain data, as it is the main purpose of the application. Other fields may be left empty.

2.4 Volpiano

The melodies in the volpiano fields are encoded as strings of alphanumeric characters and dashes. These can be rendered as musical notation using the volpiano font. Each character represents either a pitch, empty space, or other musical characters, such as a clef.

Volpiano was developed as a research tool optimized for databases and word processors. There are strict rules concerning the transcription, which leads to all volpiano-encoded melodies having a standardized format. Each transcription begins with a treble clef. Gaps between words are encoded as three dashes, while two dashes represent gaps between syllables (Helsen and Lacoste [2011]).

2.5 Data cleaning

Conclusion

Bibliography

Olivier Berten and contributors. Gregobase: A database of gregorian scores. URL https://gregobase.selapa.net/.

Dimitrios Bountouridis, Daniel G. Brown, Frans Wiering, and Remco C. Veltkamp. Melodic similarity and applications using biologically-inspired techniques. *Applied Sciences*, 7(12), 2017. ISSN 2076-3417. doi: 10.3390/app7121242. URL https://www.mdpi.com/2076-3417/7/12/1242.

Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne. Studying large plainchant corpora using chant21. In *7th International Conference on Digital Libraries for Musicology*, DLfM 2020, page 40–44, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450387606. doi: 10.1145/3424911.3425514. URL https://doi.org/10.1145/3424911.3425514.

Bas Cornelissen, Willen Zuidema, and John Ashley Burgoyne. Mode classification and natural units in plainchant. 2020b. URL https://program.ismir2020.org/poster_232.html.

Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. pages 637–642, 2010.

Kate Helsen and Debra Lacoste. A report on the encoding of melodic incipits in the cantus database with the music font 'volpiano'. *Plainsong and Medieval Music*, 20(1):51–65, 2011. doi: 10.1017/S0961137110000197. URL https://doi.org/10.1017/S0961137110000197.

David Hiley. *Gregorian Chant*. Cambridge Introductions to Music. Cambridge University Press, 2009. doi: 10.1017/CBO9780511807848.

David Huron. The melodic arch in western folksongs. *Computing in Musicology*, pages 3–23, 1996.

David Huron and Joshua Veltman. A cognitive approach to medieval mode: Evidence for an historical antecedent to the major/minor system. *Empirical Musicology Review*, 1(1):33–55, 2006. URL https://doi.org/10.18061/1811/24072.

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

Debra Lacoste. The cantus database: Mining for medieval chant traditions. *Digital Medievalist*, 7, 2012. URL http://doi.org/10.16995/dm.42.

Debra Lacoste and Jan Koláček. Cantus index: Online catalog for mass and office chants. URL http://cantusindex.org/.

Debra Lacoste, Jan Koláček, Terence Bailey, and Ruth Steiner. A database for latin ecclesiastical chant - inventories of chant sources. URL https://cantus.uwaterloo.ca/.

Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(70)90057-4. URL https://www.sciencedirect.com/science/article/pii/0022283670900574.

Cédric Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3, 2007. URL https://doi.org/10.1371/journal.pcbi.0030123.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: a novel method for fast and accurate multiple sequence alignment11edited by j. thornton. *Journal of Molecular Biology*, 302(1):205–217, 2000. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.2000.4042. URL https://www.sciencedirect.com/science/article/pii/S0022283600940427.

Saebyul Park, Taegyun Kwon, Jongpil Lee, Jeounghoon Kim, and Juhan Nam. A cross-scape plot representation for visualizing symbolic melodic similarity. pages 423–430, 2019.

Rebecca Shaw. Differentiae in the cantus manuscript database: Standardization and musicological application. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, DLfM '18, page 38–46, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365222. doi: 10.1145/3273024.3273028. URL https://doi.org/10.1145/3273024.3273028.

Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1:337–348, 2009. URL http://doi.org/10.1089/cmb.1994.1.337.

David D. Wickland. Evaluating melodic similarity using pairwise sequence alignments and suffix trees. Master's thesis, The University of Guelph, 9 2017.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment