MASTERARBEIT | MASTER'S THESIS

Titel | Title

Extrapolating the relationship between folk music melodies based on musical patterns

verfasst von | submitted by

Vanessa Veitsch BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 878
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Verhaltens-, Neuro- und Kognitionsbiologie
Betreut von | Supervisor:	Univ.-Prof. William Tecumseh Sherman Fitch PhD

Acknowledgements

At point of my master's thesis I want to express my gratitude to a few special people. First of all, I want to thank my supervisor Univ.-Prof. William Tecumseh Sherman Fitch, for giving me the opportunity to work on such an interesting and unlikely topic, even if in the end it didn't work completely as planned and several obstacles had to be overcome for this thesis. Further I want to thank him for his much-appreciated feedback and patience.

Furthermore, I want to thank Tecumseh's research group for valuable discussions and input. A special thanks hereby goes to Jeroen Van der Aa, who tried his best to assist me in one addition to this master's thesis, which unfortunately had to be abandoned due to its complexity.

Last but not least, I want to offer my heartfelt thanks to Martin Mauser, who always supported me with advice and action, as well as my family and friends.

Abstract

This thesis investigates the structural and evolutionary relationship among English folk music melodies using computational methods which were inspired by the field of bioinformatics. Hereby genetic sequence alignment was applied to musical evolution in the special case of 1198 songs from Volume 2 of Bronson Child Ballads. The melodies were transcribed, verified and standardized by transposing them into C major or a minor using the Krumhansl-Schmuckler key-finding algorithm. Melodic similarity was assessed using the Levenshtein edit distance, while Shannon entropy was employed to evaluate rhythmic and pitch complexity. The obtained results indicate that songs within the same lyrical chapter exhibit greater melodic similarity than those across chapters. However, entropy analyses revealed no significant variation in rhythmic or pitch predictability between chapters. These findings suggest that while lyrical grouping may imply melodic relations, it cannot be guaranteed. The study highlights the potential of interdisciplinary approaches, particularly those borrowed from genetics and information theory, to investigate patterns in musical evolution and cultural transmission.

Keywords:
Bronson Child Ballads, Bioinformatics, Melody analysis, Information theory, Musical evolution

Zusammenfassung

Diese Arbeit untersucht die strukturellen und evolutionären Beziehungen zwischen englischen Volksmusikmelodien unter Verwendung computergestützter Methoden, die aus dem Bereich der Bioinformatik inspiriert sind. Dabei wurde die genetische Sequenzalignment auf die musikalische Evolution im speziellen Fall von 1198 Liedern aus Band 2 der Bronson Child Balladen angewendet. Die Melodien wurden transkribiert, verifiziert und standardisiert, indem sie mit Hilfe des Krumhansl-Schmuckler-Algorithmus zur Tonartbestimmung in C-Dur oder a-Moll transponiert wurden. Die melodische Ähnlichkeit wurde anhand der Levenshtein edit distance bewertet, während die rhythmische und tonhöhenbezogene Komplexität mit Hilfe der Shannon-Entropie bewertet wurde. Die erzielten Ergebnisse zeigen, dass Lieder innerhalb desselben lyrischen Kapitels eine größere melodische Ähnlichkeit aufweisen als Lieder aus verschiedenen Kapiteln. Die Entropieanalysen ergaben jedoch keine signifikanten Unterschiede in der Vorhersagbarkeit von Rhythmus und Tonhöhe zwischen den Kapiteln. Diese Ergebnisse deuten darauf hin, dass eine lyrische Gruppierung zwar melodische Beziehungen implizieren kann, diese jedoch nicht garantiert sind. Die Studie unterstreicht das Potenzial interdisziplinärer Ansätze, insbesondere solcher, die aus der Genetik und der Informationstheorie übernommen wurden, um Muster in der musikalischen Evolution und kulturellen Übertragung zu untersuchen.

Schlüsselwörter:
Bronson Child Balladen, Bioinformatik, melodische Analysen, Informationstheorie, musikalische Evolution

Table of contents

1. Introduction

In 2022, Savage et al investigated the evolution of folk song melodies from Japan and England by identifying cross-cultural regularities in the evolution of music with help of genetical mechanisms. That novel work leveraged the identification that music is following predictable rules. One key observation was that notes are less common to change if they show stronger rhythmic functions. Further could be shown that substitutions are less likely than note insertions and deletions, also known as indels (Savage et al., 2022). This mechanism describes one of many in the field of genetics. Hereby insertions mean that nucleotide base pairs are added or in case of deletion removed from the DNA molecule (Herrera and Garcia-Bertrand, 2023).

Another key finding was that substitutions, which describes the process of replacement in genetics, occur most of the time between neighbouring notes, indicating that smaller note changes are more persistent than larger derivations. Based on the identified connection between genetics and musical evolution the authors used sequence alignment algorithms from the field of molecular biology to investigate the evolution of music. Equipped with that tool 10062 songs were analysed, identifying pairs of 328 close related melodies which were subsequently investigated in relation to their patterns of note changes.
Based on the conducted analysis the final and the stressed notes, which are more pronounced than other ones, tend to be more stable since they are notes with stronger rhythmic roles (Zimmerman, 1997). On the opposite notes which are less influential for the rhythm, e.g. ornamental notes, which are embellishments, are less stable (Bakirova, 2020). Furthermore, notes which are more narrow in pitch, which means the distance between the smallest and the highest pitch is small, show a higher occurrence of substitutions (“Melody | Music Appreciation 1,” n.d.).
To summarize the work from Patrick E. Savage et al, music shows, like genetics, underlying evolutionary constraints. Which means that cross-cultural regularities can be exhibited in cultural evolution which leads furthermore to a better understanding of change and diversity in music (Savage et al., 2022).

Within this thesis we will use these observations from Savage et al and apply them to a collection of folk tunes, which are collected in the Bronson Child Ballads. To be precise, the Bronson Child Ballads are a collection of a collection. In the second half of the 19th century Francis James Child wrote a collection of songs with the title “The

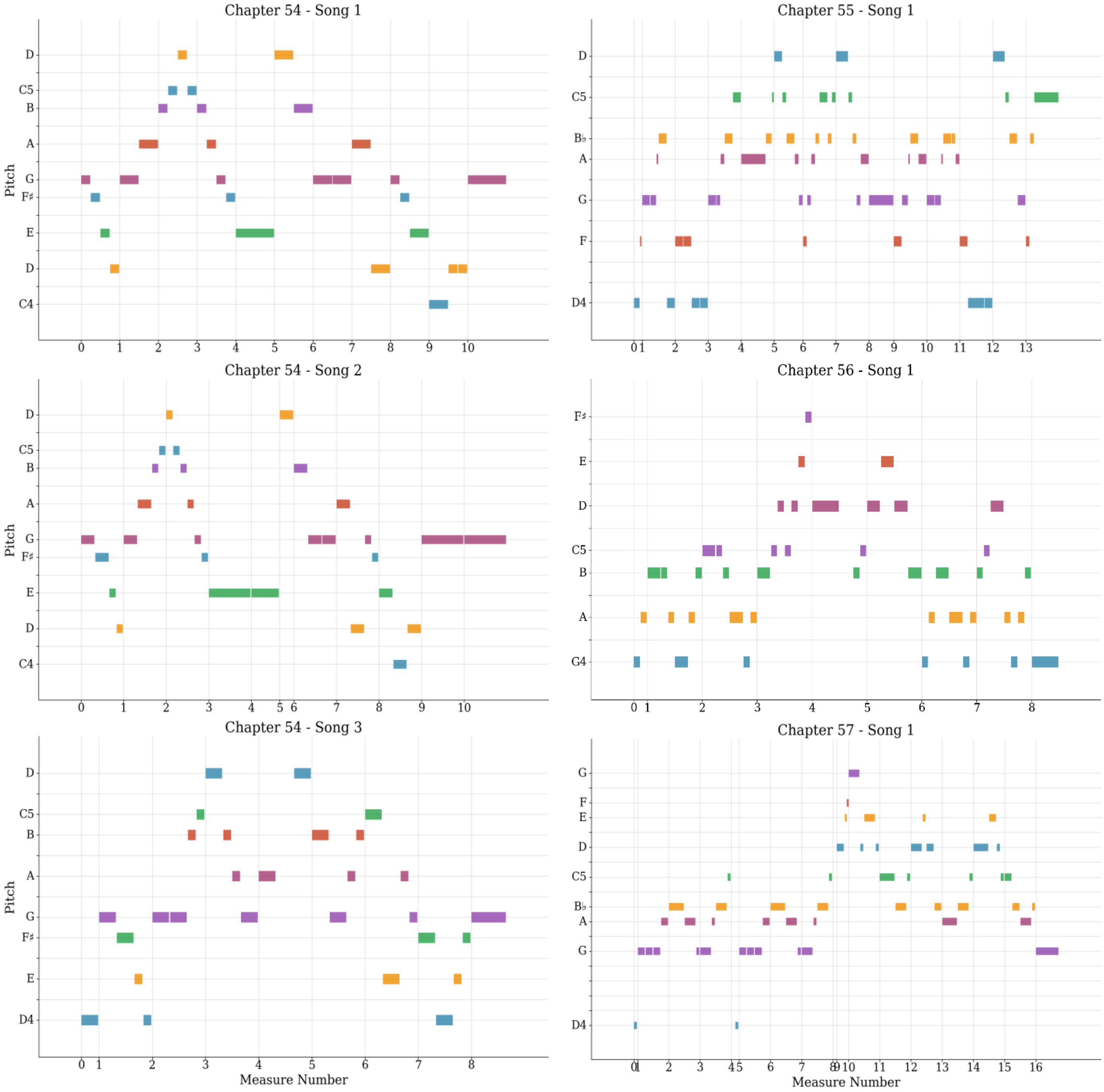
English and Scottish Popular Ballads". Up to this point in history there hasn't been such huge compilation of ballads before (Child et al., 1904). Later in the 1960ties Bertrand Harris Bronson annotated and organized the collection of Francis James Child into tune families with similar lyrics. In addition, Bronson completed the version of Child, where only texts were written down, with different folk tunes from England, Scotland, Ireland and North America. In the final iteration the collection is a compilation of four different volumes consisting out of approximately 4120 tunes (Simpson, 1960).

As a first step we should define what ballads actually are. Ballads are songs which tell a story where the author doesn't drag personal things into the story. In order to sing along ballads are rather simple with a metrical structure organized into short stanzas, which describes a reoccurring pattern of rhyme and meter ("OnMusic Dictionary - Term," n.d.). Furthermore, they are meant to be passed on orally and therefor undergo the same processes of change like the ordinary language, thereby rendering it the ideal dataset for this master's thesis, in order to identify the cultural relationship of Scottish, English, Irish and North American folk songs (Child et al., 1904) (Moore, 1916).

To get a better feeling of the data, the following figure shows examples of songs from the used data. Hereby three songs are from one song family or chapter and three from different song families or chapters as one can see in figure 1 below.

But why is this master's thesis about music written in a biological field? What is the relationship between music and biology? Music shows a great diversity and is the result of musicality which is strongly connected to our biology. Therefore, musicality can be studied from many perspectives like neural, cognitive, developmental and comparative ones (Fitch, 2015).

As previously mentioned, the Bronson Child Ballads are a suitable testing bed to investigate the evolution of music. Since the idea of Savage et al was to reveal cross cultural regularities, this thesis wants to investigate the relationship between and within song families from the Bronson Child ballads, more precisely Volume 2. Hereby the assumption was made that songs in the same chapter or song family are closely related, while songs in different chapters show less correlation. Thereby we are assuming that songs which share the same lyrics are closer related to those with different lyrics. This research question will be evaluated in the next few chapters starting with Chapter 2 “Material and Methods”.


Chapter 54 - Song 1

Chapter 55 - Song 1

Chapter 54 - Song 2

Chapter 56 - Song 1

Chapter 54 - Song 3

Chapter 57 - Song 1

Figure 1: This figure represents the piano rolls from different songs, providing a rough overview of the songs within a song family and from different ones. The three songs in the left column are the first three songs from chapter 54. On the right side are three songs from different chapters as followed: the first one is the first from chapter 55, the second one is the first from chapter 56 and the third song is the first one from chapter 57. Each graph shows on the x-axis the measure number over time, i.e. the length in quarter notes. On the y-axis the pitch from each note.

2. Material & Methods

2.1. The data

The raw data was taken from Bronson Child Ballads, consisting out of four Volumes. A former project treated the data which it has been partially transcribed into the ABC format. EasyABC is an open source editor for music notation, which transcribes music into the ABC format, which was originally developed by Nils Liberg in 2014 (Wybren de Jong and Shlien, 2021).

Due to completeness of the beforehand taken transcription of the songs and personal preference the second Volume was used for the remainder of this thesis. Volume 2 consists of approximately 1198 songs which were checked, in sight of correctness of the transcription, by hand. The process of verifying the rightness of the data included partially listening to songs and checking each note by reading simultaneously the musical notation and the annotated version. Since only the first paragraph of the song text from every transcribed song was notated, the missing ones were added from the original data. However, due to time constraints not all songs got checked according to their lyrics which limits further analysis to the notes. After correcting those files new subfolders were created, one for the complete Vol 2 where the melodic part was checked and one for the lyrics where just a few chapters or song families were checked.

2.2. Used programs

For the transcription part of this thesis, the already mentioned opensource software EasyABC was used. Furthermore, the programming language Python in the Jupyterlab environment was utilized. For the musical analysis of the different songs the package Music21 was used.

Table 1 shows a list of the used programs and their versions, in order to enable verification of the presented results.

Table 1: Versions of the used libraries.

python	3.10.12
jupyterlab	4.0.6
matplotlib	3.8.0
music21	9.1.0
musicdiff	4.0.0
numpy	1.26.0
pandas	2.1.1
scipy	1.11.3
seaborn	0.13.0
tqdm	4.66.2

2.3. Analysis of the raw data

To get a better understanding of the different songs, a small analysis was done beforehand. Here in the duration of the different songs, the pitch range and the median height of pitches were measured. Additionally, an overview of the used keys, which describes the organizational base of a song, within one song family was prepared (“Key | Music, Major, Minor, & Chromatic Scales | Britannica,” n.d.).

2.4. Preparation of the data for further analyses

All songs are grouped into different song families, which are in the folder described as different chapters in the Volume. To get a better comparison within and between those families all songs were translated to either C major or a minor. To achieve the transposition a key-finding algorithm, which was developed by Carol Krumhansl and Mark A. Schmuckler, was used. This is embedded into the transpose function in Music21, it looks for the most likely key of a melody. Therefore, 24 pitch classes, since we have major and minor, were represented as key profiles which are based on listener ratings from prior psychological experiments. The algorithm calculates the total duration of each pitch class in the original music and then computes the Pearson correlation between this data and each rotated key profile corresponding to the above-mentioned 24 keys. In the end the key with the highest correlation is selected (Hart, 2012).

After that, a pickle file was created as easily accessible file which contains the title, original key, length (in beats), range and the melody itself.

The following two graphs show the reliability of the transposed songs. This was tested with the visualization of Krumhansl-Schmuckler Key Analysis. As one can see, the left graph of figure 2 shows the song with the original key (A major). On the right side the transposed version of the same song in C major is shown.

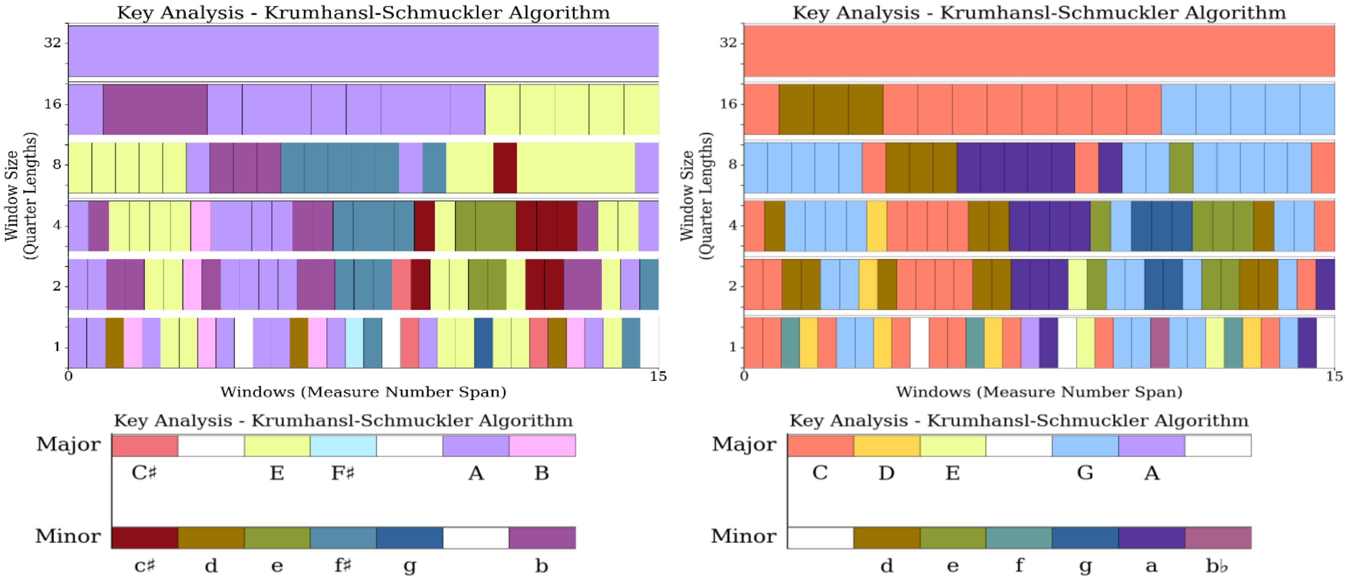


Figure 2: This figure shows the first song of the data prior and after the transposition. The left image shows the non-transposed song and the right image the transposed version. On the bottom of each graph is the legend to show which notes were used. Hereby the smaller the window size the more accurate is the representation of the individual notes.

2.5. Levenshtein edit distance

Another key goal in this thesis was to measure the difference among and between different song families. Therefore, the Levenshtein edit distance was used to reach this goal. The Levenshtein edit distance is a string similarity metric which means that it measures the minimum number of single-character edits. Singe character edits are deletion, insertion and substitution. Each cell represents the edit distance between the first character *i* in the source string and character *j* in the target string. The final cell gives the total edit distance of the required numbers of single-character edits from *i* to j (Haldar and Mukhopadhyay, n.d.). The paper “A diff procedure for music score files” by Francesco Foscarin et al. was used as the basis for this study. In it he proposed a method to compare two score files at the level of notation (Foscarin et al., 2019).

2.6. Shannon entropy

Another aspect of this work is to compare the Shannon entropy within the chapters. The Shannon entropy, which is used for binary systems, is described with the following formula:

$$H(X) \coloneqq -\sum_{x \in X} p(x) \log_p(x)$$

where $H(X)$ describes the Sannon entropy depending on the dataset X, while $p(x)$ represents the probability of occurrence of x.

With this formula you can reach a minimum of 0 and a maximum of 1 conditioned on binary values, whereas 0 stands for the most certain situation and 1 for the most uncertain which can occur (Shannon, 1948).

With this formula it was tried to calculate the Sannon entropy of different note values and pitch values. Whereas note values describe the relative durations of a note and pitch values the height of a tone in respect to a musical scale as in this case the C scale. To accomplish this point in this thesis the paper from Mihail Miller about “Vergleichende Analyse der Komplexität musikalischer Dimensionen von Klavierwerken europäischer Komponisten anhand von MIDI-Files” was used (Miller, 2020). In it she describes the complexity of piano works from European composers like F. Chopin and J.S. Bach. The author investigated pitch class, inter-onset interval and melodic interval with the methods from (Conklin, 2006) and (Madsen and Widmer, n.d.).

3. Results

3.1. General analysis

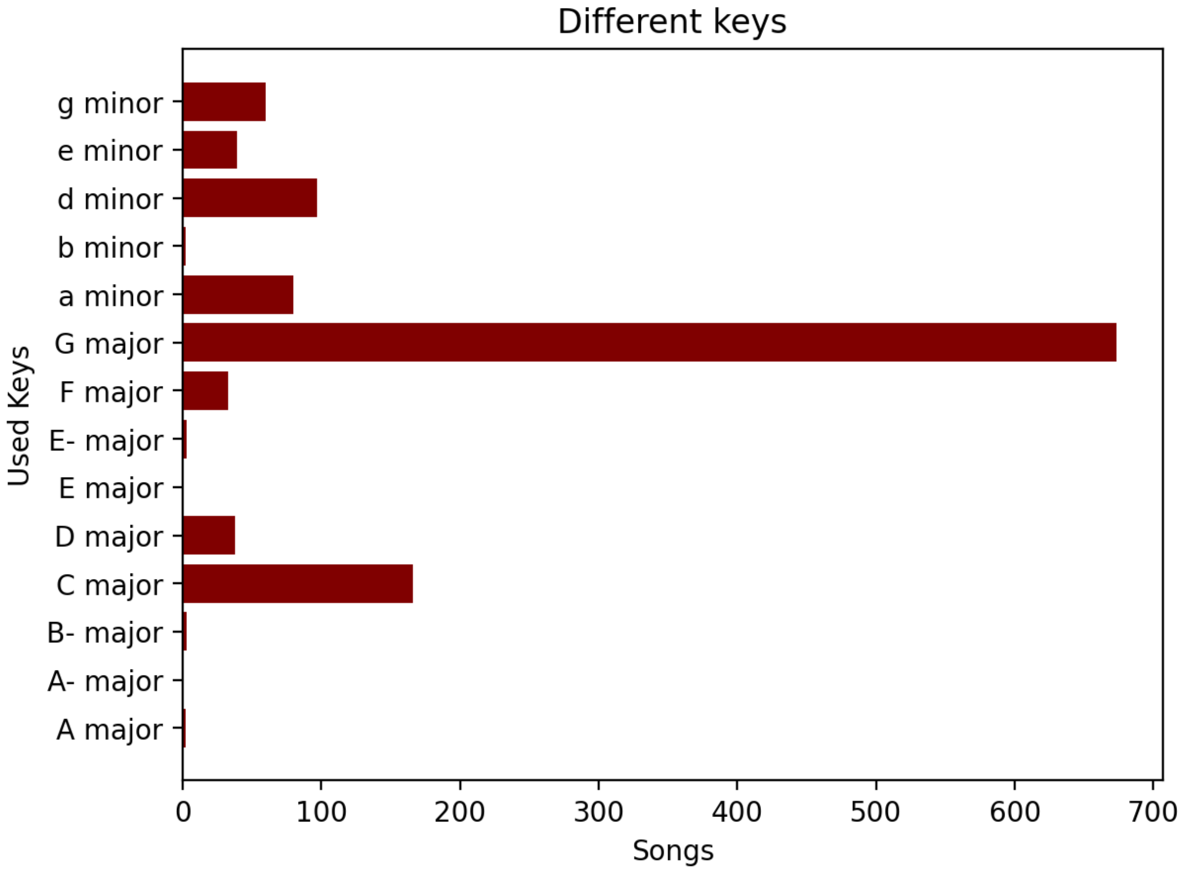


Figure 3: This figure shows the distribution of different keys which were used in different songs. On the left side one can see the keys used in major and minor and on the bottom the number of songs are represented.

Figure 3 gives an overview of the different keys which were used in this master's thesis of Volume 2 in Bronson Child Ballads. One can identify a wide range of keys where the most common key is G major. The second most common key is C major followed by d minor. Keys A major and E major were the least used.

Furthermore, the variation of the length of the various songs, were analysed and visualized in figure 4 provides more detailed information. As one can see in figure 4 the majority of songs, about 155, have a length of 29 beats. (The length of a beat is determined by the tempo of the songs. Since all songs have a tempo of 120 it means that each second 2 beats occur.) The second most common length of beats lies approximately by 46 beats.

Further statistical tests were conducted to get the mean, minimum, maximum, variance, skewness and kurtosis of the median heights. The results are listed in table 2.

Table 2: Statistic of the lengths of different songs

minimum	15.0
maximum	128.0
mean	39.05
variance	256.50
skewness	1.67
kurtosis	4.23

Herewith, the skewness γ – or 3rd central motion – is defined as

$$\gamma \;= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})^3}{\sigma^3}\right]$$

Where N defines the ideal number of samples, X_i describes the individual samples, and $\bar{X}$ and σ describe the mean and standard deviation, respectively. Is γ > 0, the distribution is right skewed and if γ < 0, it is skewed to the left. So, if we look at the results from table 2 one can see the skewness is above 0, indicating that in the dataset there is a long tail to the right consisting of songs with much longer length than the mean value.

The kurtosis is defined as

$$\omega = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathcal{X}_i - \bar{x}}{\sigma}\right)^4$$

hereby $\mathcal{X}_i$ defines the observation values, $\bar{x}$ the arithmetic mean and σ the standard deviation. Herewith, if ω >3 the distribution is peaked and if ω < 3 the distribution is flat-top. Based on the results of table 2 the kurtosis is 4.23, which means the distribution is peaked.

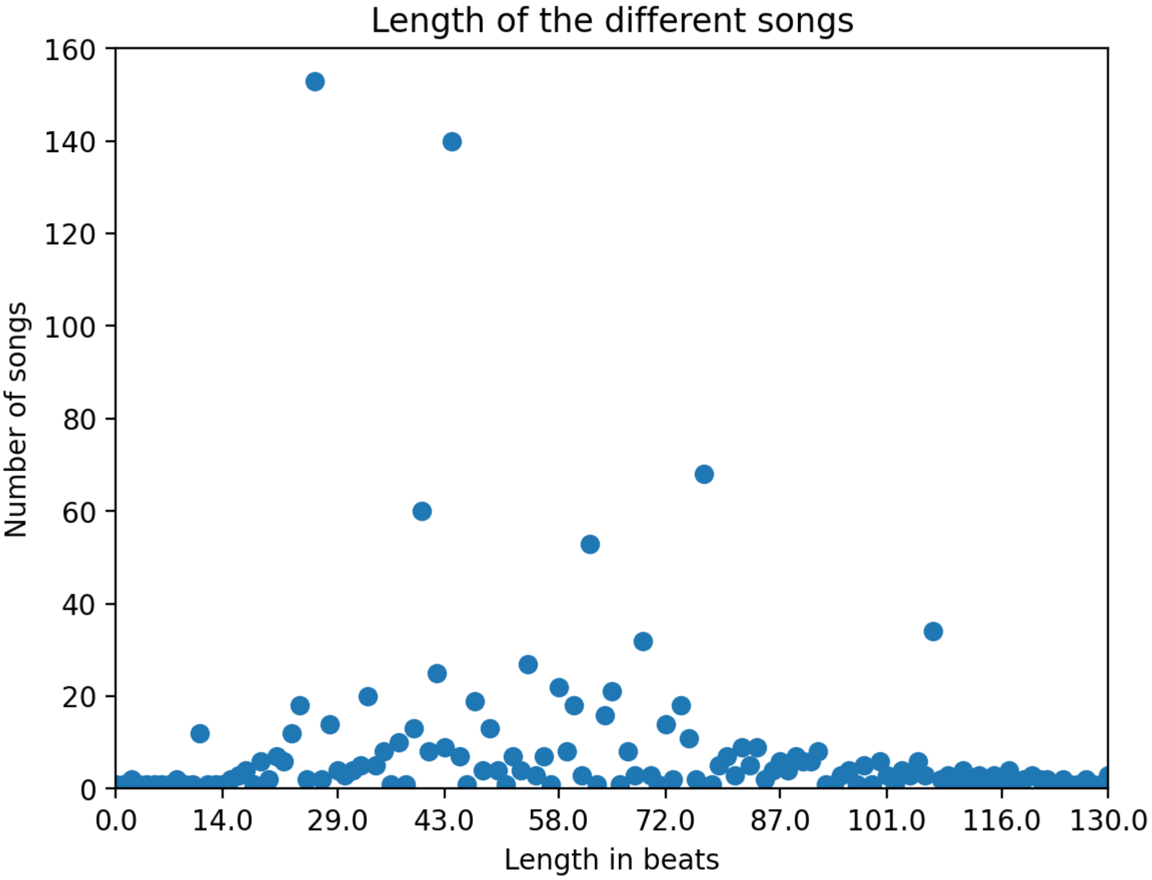


Figure 4: This figure shows the different lengths of songs across all songs in Volume 2. On the left side of the figure are the total numbers of songs according to the length measured in beats on the bottom.

Figure 5 visualizes the different median heights. One can see that the highest number of songs show a medium height of 69.0. It is important to keep in mind, that these numerical values correspond to the MIDI values. Herewith a middle C (C4) corresponds to a value of 60. Table 3 summarizes the statistical key points of the mean heights of all songs within the examined chapter. In order to ease conversion between MIDI values and written notes figure 6 (Wolfe, n.d.) can be consulted. The most common value, namely 69 with around 400 songs, corresponds to an A4. The second most common value is 67which corresponds to G4. On the third place with around 200 songs is the median height of 71.0 or according to table 2 a B4. The least common median heights are 63.0 (D#4 od Eb4) and 62.0 (D4).

As a side information for better understanding the MIDI, also known as Musical Instrument Digital Interface, value describes a standardization to enable a transmission between electronic instruments and computers of digitally encoded music information which are sound synthesizing capable. With it, it is possible to process key, pitch and velocity of a note and perform statistical tests with it (Salmon and Newmark, 1989).

Table 3 visualizes the statistic of the different median heights. This statistic uses the same formulas which were already described beforehand in table 2. The highest height was 74.0 and the lowest was 62.0. Here we have as well a right skewed distribution with 0.20 but a flat-top kurtosis with -0.11.

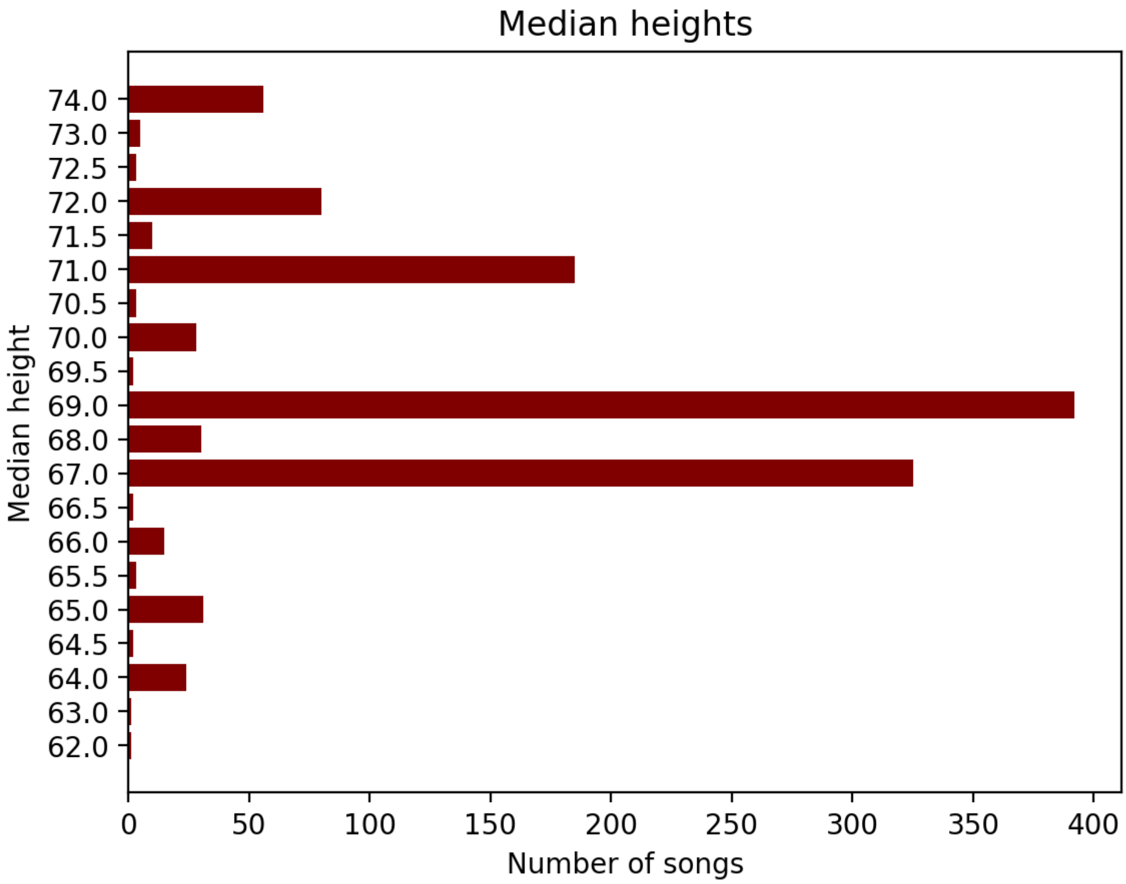
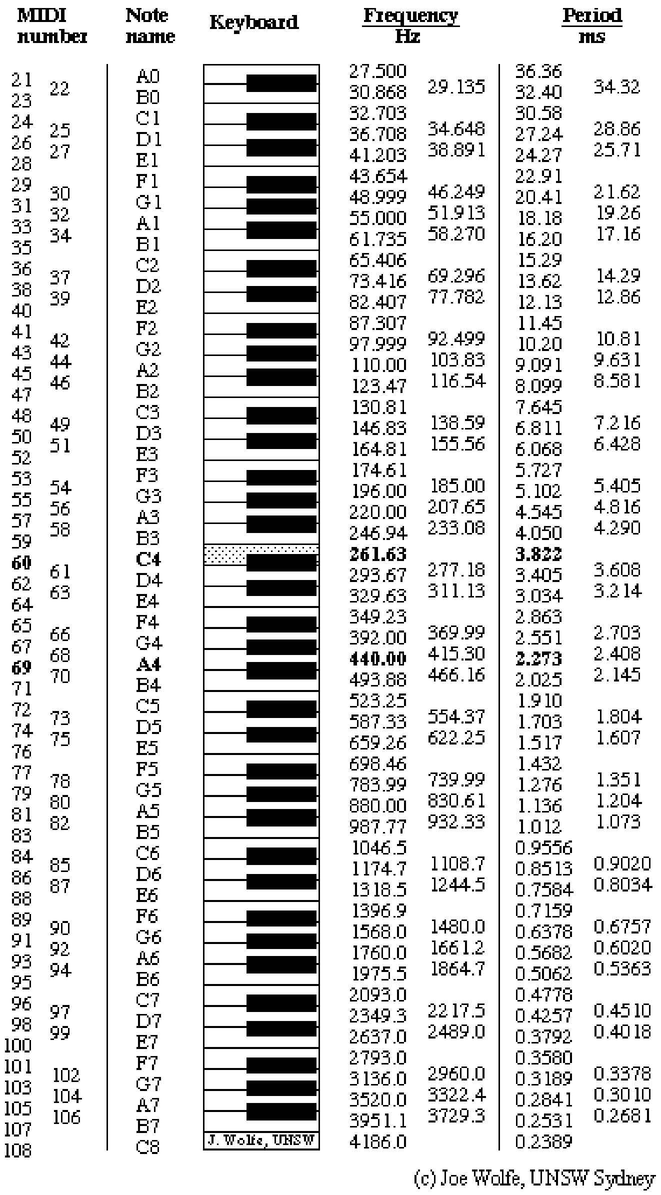


Figure 5: The figure represents the median height of all songs in Volume 2. On the y-axis one can see the median height, where the middle C is equal to 60. On the x-axis are the total number of songs which are shown as bars corresponding to the median height.

Table 3: Statistic of the median heights

minimum	62.0
maximum	74.0
mean	69.0
variance	4.83
skewness	0.20
kurtosis	-0.11



MIDI number	Note name	Frequency Hz	Period ms
21	A0	27.500	36.36
22		29.135	34.32
23	B0	30.868	32.40
24	C1	32.703	30.58
25		34.648	28.86
26	D1	36.708	27.24
27		38.891	25.71
28	E1	41.203	24.27
29	F1	43.654	22.91
30		46.249	21.62
31	G1	48.999	20.41
32		51.913	19.26
33	A1	55.000	18.18
34		58.270	17.16
35	B1	61.735	16.20
36	C2	65.406	15.29
37		69.296	14.29
38	D2	73.416	13.62
39		77.782	12.86
40	E2	82.407	12.13
41	F2	87.307	11.45
42		92.499	10.81
43	G2	97.999	10.20
44		103.83	9.631
45	A2	110.00	9.091
46		116.54	8.581
47	B2	123.47	8.099
48	C3	130.81	7.645
49		138.59	7.216
50	D3	146.83	6.811
51		155.56	6.428
52	E3	164.81	6.068
53	F3	174.61	5.727
54		185.00	5.405
55	G3	196.00	5.102
56		207.65	4.816
57	A3	220.00	4.545
58		233.08	4.290
59	B3	246.94	4.050
60	**C4**	**261.63**	**3.822**
61		277.18	3.608
62	D4	293.67	3.405
63		311.13	3.214
64	E4	329.63	3.034
65	F4	349.23	2.863
66		369.99	2.703
67	G4	392.00	2.551
68		415.30	2.408
69	**A4**	**440.00**	**2.273**
70		466.16	2.145
71	B4	493.88	2.025
72	C5	523.25	1.910
73		554.37	1.804
74	D5	587.33	1.703
75		622.25	1.607
76	E5	659.26	1.517
77	F5	698.46	1.432
78		739.99	1.351
79	G5	783.99	1.276
80		830.61	1.204
81	A5	880.00	1.136
82		932.33	1.073
83	B5	987.77	1.012
84	C6	1046.5	0.9556
85		1108.7	0.9020
86	D6	1174.7	0.8513
87		1244.5	0.8034
88	E6	1318.5	0.7584
89	F6	1396.9	0.7159
90		1480.0	0.6757
91	G6	1568.0	0.6378
92		1661.2	0.6020
93	A6	1760.0	0.5682
94		1864.7	0.5363
95	B6	1975.5	0.5062
96	C7	2093.0	0.4778
97		2217.5	0.4510
98	D7	2349.3	0.4257
99		2489.0	0.4018
100	E7	2637.0	0.3792
101	F7	2793.0	0.3580
102		2960.0	0.3378
103	G7	3136.0	0.3189
104		3322.4	0.3010
105	A7	3520.0	0.2841
106		3729.3	0.2681
107	B7	3951.1	0.2531
108	C8	4186.0	0.2389

(c) Joe Wolfe, UNSW Sydney

Figure 6: This table visualizes the MIDI numbers to their corresponding note names, frequencies and periods (Wolfe, n.d.).

3.2. Levenshtein edit distance

Figure 7 represents the mean of the Levenshtein edit distance. Typically, this is used to compare the DNA sequences of proteins but here it is used to calculate the differences between notes within a song family and between song families, bringing the connection between genetics and musical evolution. On the diagonal songs of the same song families are shown. Hereby, one can identify five black squares. Those are the families which have just one song in it, therefore they have an edit distance of 0 or in other words they are identical. However, in general no clear similarity can be drawn visually between the different song families nor within the song families. One thing which can be seen is that the 30th, 32nd, 34th, 38th and the 39th chapter show a low level of agreement with the remainder of the dataset because they show an edit distance between 70 and 80 (figure 6).

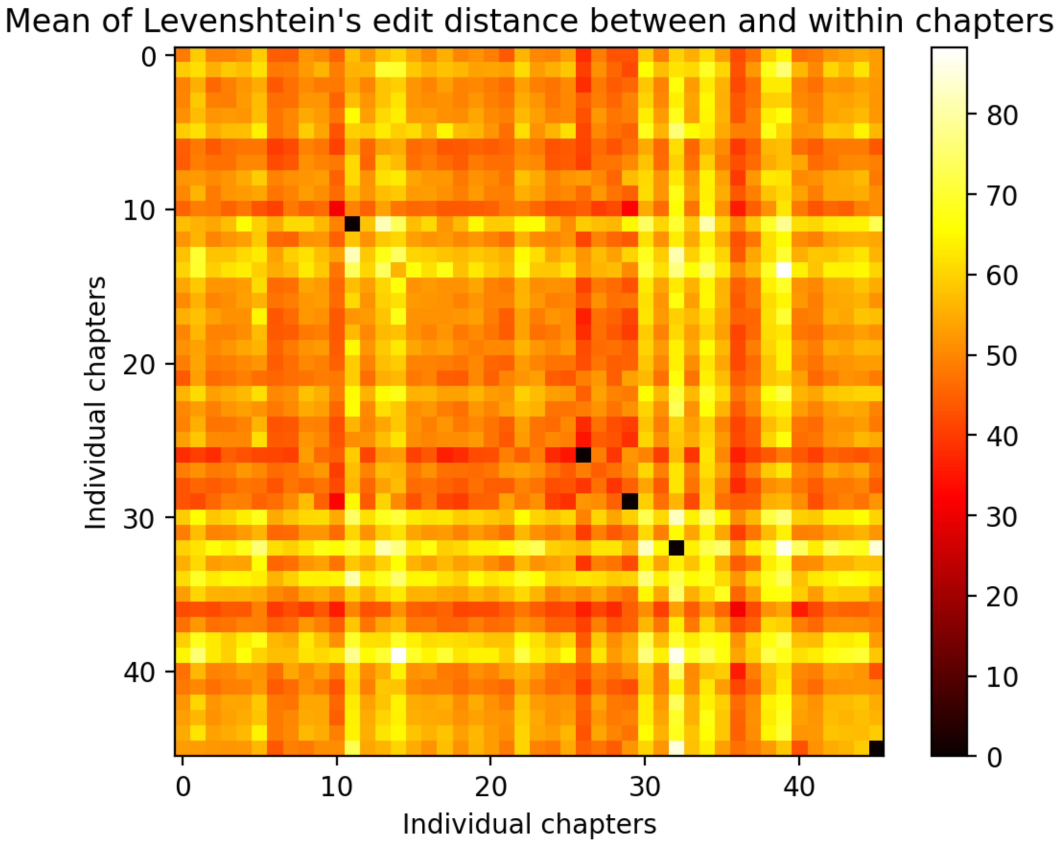


Figure 7: The figure represents the average Levenshtein edit distance between different chapters and within the chapters. On the x- and y- axis are the individual chapters shown. On the right side of this figure one can see a scale which shows the differences in colour. The darker the colour is, the smaller is the difference between two data points. On the diagonal are the differences within the chapters whereas the rest shows the similarity matrix for the differences between the different chapters which are symmetric around the diagonal. All songs were converted into C major or a minor to ensure correctness of Levenshtein edit distance before analysis.

Table 4: gives an overview of the minimal and maximal Levenshtein edit distance of the individual song pairings. Hereby the song numbers and the corresponding chapters are given. On the top of the table are the closest related songs and on the bottom the most distant.

Song	**Chapter**	**Song**	**Chapter**	**Edit distance**
730	81	731	81	1
868	84	869	84	12
827	84	849	84	13
509	75	510	75	14
311	73	318	73	14
270	73	1103	95	271
256	68	270	73	271
270	73	1119	95	273
270	73	995	85	276
270	73	443	74	283

An overview about the mean of all calculated edit distances which is 51.83, the smallest calculated value 1.0 and the highest value 283.0 was done. Furthermore, table 4 shows the most similar and the most different song pairings. For example, the most similar songs are song 730 and 731 both from chapter 81 with an edit distance of 1.0. The most different songs are song 270 from chapter 73 and song 443 from chapter 74 with an edit distance of 283. What else can be seen in this table is that similar songs tend more to be in the same chapter or song family.

In order to quantify this assumption an edit distance dependent measure was utilized. To achieve this following formula was used:

$$p = \frac{\#song\ pairings \,\big|\, edit\ distance < mean}{\#song\ pairings}$$

Hereby $\#song\ pairings$ can be used for values within the same or between different chapters. Further p gives the relative probability of a song pairing (either within the same or in between different chapters) to have a lower edit distance than the mean edit distance of all possible song pairings (both within and in between chapters).The outcome was that song pairings, within the same chapter, have a probability of ~61.03% that they obey a smaller edit distance than the average edit distance of all song pairings.
In contrast song pairings in between different chapters only have a smaller than average edit distance with a probability of ~55.06%.
Thereby we can say that it is more likely for songs in the same chapter to have a small edit distance.

When it comes to the statistical part, a one way ANOVA was performed. The outcome was significant since the p-value is 0.0 and the F-statistic 88.09. Since the significance level is <0.05, the Null-hypothesis can be rejected and we can say at least one song pairing is significantly different from the others.

Further if one has a look on the standard deviation of Levenshtein edit distance, which is shown in figure 8, it is noticeable that the standard deviation is overall quite small. Which shows that the data points are close together and don’t deviate far from the mean value. Two noticeable outliers are the song pairings between chapter 11 and 14, which obey a standard deviation higher than the mean edit distance of the entire data set.

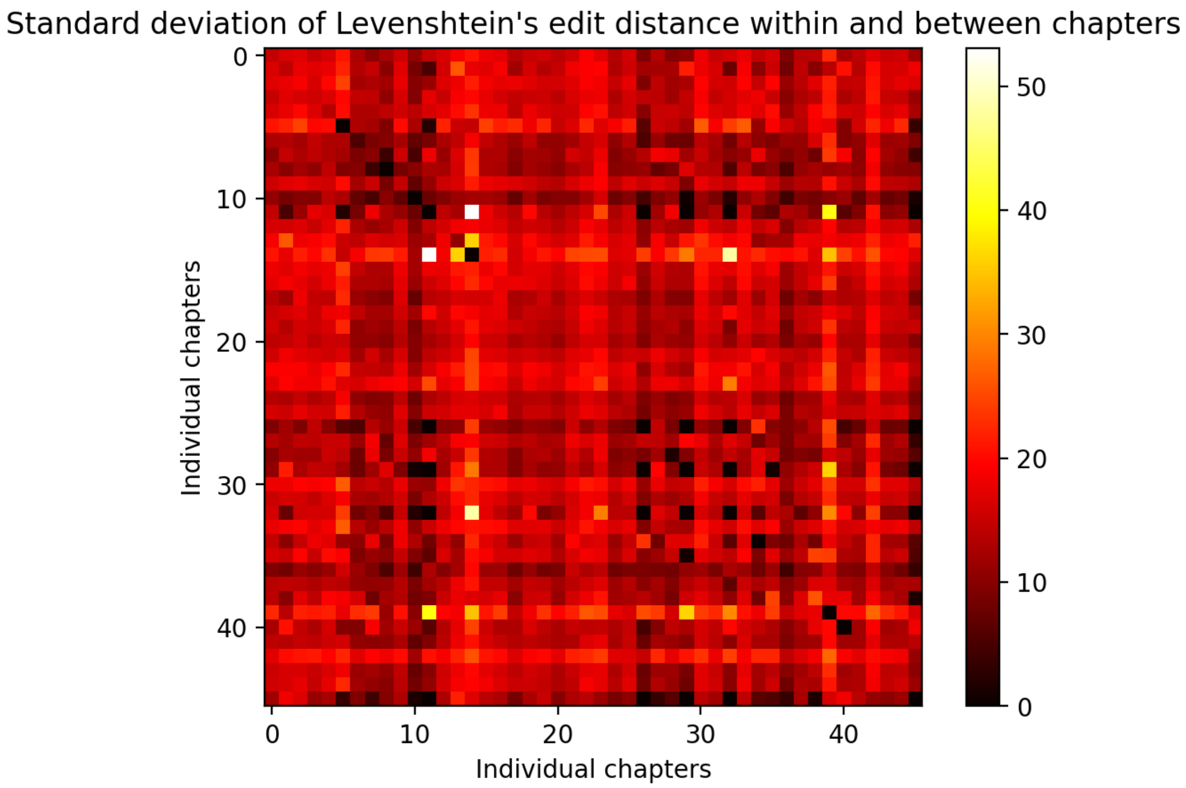


Figure 8: This figure represents the standard deviation between different chapters (off diagonal) and within the chapters (diagonal). On the x- and y- axis are the individual chapters shown. On the right side of this figure one can see a scale which shows the standard deviation. The darker the colour is the smaller standard deviation between two chapters.

3.3. Shannon entropy

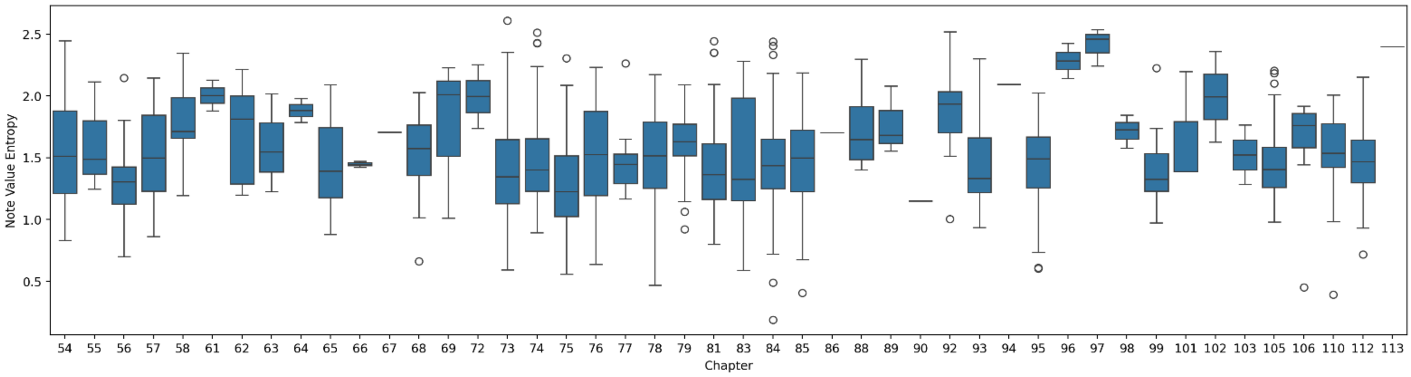


Figure 9: This figure visualizes the calculated Shannon entropy of note values within the same chapters. On the x-axis are the different chapters shown and on the y-axis the note value entropy. The songs were beforehand put into C major and a minor.

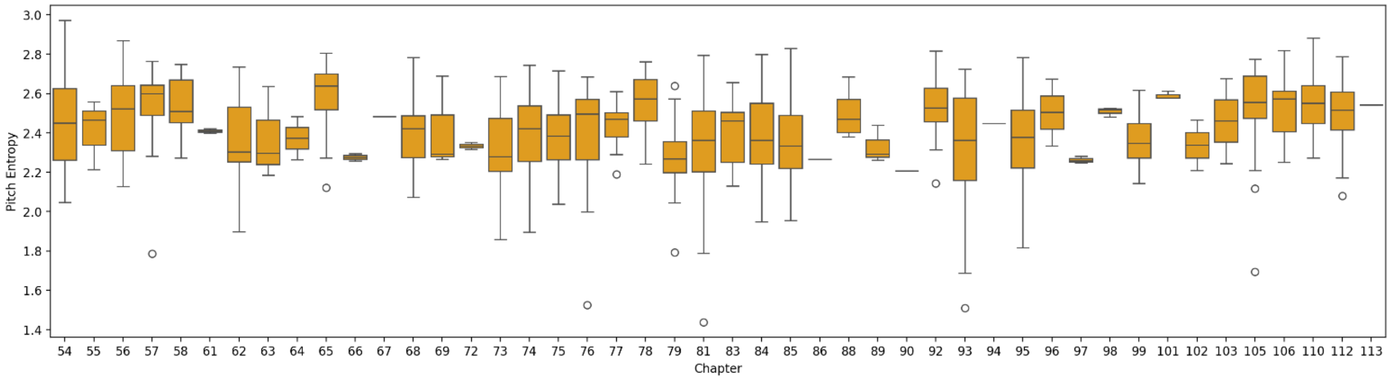


Figure 10: This figure represents the Shannon entropy of pitch values. On the x-axis are the different chapters shown and on the y-axis are the matching results of the pitch entropy plotted. The songs were beforehand put into C major and a minor.

Figure 9 and 10 show the calculated Shannon entropy. Figure 9 describes the note value entropy. Hereby the higher the note value entropy the more uncertain it is to predict the following note values. In general, small fluctuations can be observed in figure 9. Except the chapters 96 and 97 are higher than the rest of the note value entropies. The majority of values are between 0.5 and 2.5. These values differ from the before mentioned boundaries of the Shannon entropy, since we are not working with a binary system. The performed F-statistic with an one-way ANOVA shows a value of around 3.36 with an p-value of 2.14. Therefor we have no significant difference between the different chapters regarding the note value entropy.

Considering figure 10, which represents the pitch entropy, one can see that the entropy values are higher than in figure 9. The values range between 1.6 and 3.0, which means that pitch entropy is more uncertain to predict than the note values entropy. However, figure 10 generally seems to be subject to less fluctuation than figure 9. Performing the F-Statistic with an one-way ANOVA a value of 4.11 and a p-value of 4.10 was achieved. As well as in the previously performed note value entropy a significance level couldn't be achieved.

4. Discussion

The scope of this master’s thesis was the investigation of structural and evolutionary patterns in folk music melodies. A curated dataset from Bronson Child Ballads, specifically Volume 2, containing 1198 songs, which has been transcribed and manually verified. The scientific approach was inspired by an earlier work by Savage et al, in which he and his team utilized tools from genetic methodologies to investigate melodic relationships (Savage et al., 2022).

Comparing to that earlier work of Savage et al a genetical tool like Levenshtein edit distance, which is typically found in bioinformatics and natural language processing. Furthermore, a part from general analytics like distribution of musical parameters, also a tool from information theory i.e. the Shannon entropy was utilized to gather new insights into the Bronson Child Ballads (Haldar and Mukhopadhyay, n.d.; Shannon, 1948).

In the analysis process it is important to transpose into either C major or a minor, to ensure consistent results. To verify the correctness of the transposition the Krumhansl-Schmuckler algorithm for key analysis was performed (Hart, 2012).

The Volume is subdivided into smaller chapters each of which representing distinct song families. This forms the backbone of our analysis and is the underlying structure tackled in the research question in this master’s thesis. Hereby, we have predicted that songs in the same chapter or song family are closely related, on the measure of the performed analyses, while songs in different chapters obey less correlation.

The first step to approach this research question was to extract the title of each song together with it’s original keys and identifying characteristics, i.e. length in beats, pitch range and melody.

The statistics gathered from this first extraction of properties of each piece revealed that G major, C major and d minor were the most abundant whereas A major and E major were the least common. Regarding the song length we can identify that songs with 29 beats or 46 beats were the most frequent. Hereby statistical measures revealed a right-skewed and peaked distribution indicating that shorter songs are represented more often.

Regarding the pitch analysis the most frequent median height was MIDI value 69 (A4), followed by 67 (G4) and 71 (B4), with the least common being 62 (D4) and 63 (D#4/Eb4). These pitch distributions were slightly skewed and exhibited a flat-top kurtosis, suggesting a moderate spread around the central values.

In order to properly evaluate melodic similarities between different pieces the Levenshtein edit distance was applied across and within chapters or song families. On average the edit distance was 51.83, with the span ranging from 1 to 283. Hereby the melodic similarity tends to be slightly larger within the same chapter. This observation was confirmed by an one-way ANOVA which identified a significant and persistent difference among song pairings with a p-value of 0.0 and an F-statistic od 88.09, allowing the rejection of the null hypothesis.

In a further analysis step the Shannon entropy of both note values and pitch values was calculated. The entropy for note values ranged from 0.5 to 2.5, in contrast to the pitch entropy which spanned a range from 1.6 to 3.0. This indicates that rhythmic patterns are more predictable than pitch sequences. However, in contrast to our results from the Levenshtein edit distance an ANOVA test for both entropies did not yield significant differences between song pairings within chapters or song families.

Overall, this thesis demonstrates that even if the different folk music song families are grouped by their lyrics this doesn’t automatically yield that melodies themself show significant relationships between them. However, we have identified that songs within one chapter obey larger melodic similarities than song pairings between different chapters.

Originally, for this master’s thesis a phylogenetical ansatz was planned. However, due to a lack of time, this first approach was dropped. Nevertheless, investigating the similarity between folk tunes with phylogenetical tools would provide an interesting insight, especially setting the focus on rhythmic parameters. These so called rhythmic parameters, for example the length in beats of a song, can one find at the beginning of the analysis part in this thesis (see chapter 3) or in the paper of Savage et al in which they investigate stronger and weaker beats (Savage et al., 2022).

Another promising idea would be to include the lyrics into the analysis of relationships between songs. Hereby it would be interesting to match the lyrics to the melody. Within this thesis, consideration was also given to adapting this approach. But since there is no clear link between melody and song text in in this version of Bronson Child ballads other methods for analyzation the correlation was used. Nevertheless, one way how to deal with this problem is to search for videos or performed songs where both the melody and the lyrics occur simultaneously. Then performing a natural language processing on this to examine the similarities and sentiment (Du, 2024; Mahedero et al., 2005). The reason why a sentiment analysis would be interesting is the fact that every biological process has underlying evolutionary constraints. This further holds for language and music, which could be shown in the paper of E. Glenn Schellenberg and Christian von Scheve. In which they studied emotional cues in music, like a faster tempo and if the majority is in major signal happiness. Their conclusion was that in the last decades popular songs tend to be more sad which means they show a decrease in tempo and more minor modes (Schellenberg and von Scheve, 2012).

The used data and the code developed for this thesis can be accessed via the following GitHub page:
https://github.com/VanessaVeitsch/Bronson/tree/Master-thesis

5. Literature

Bakirova, G., 2020. The History of Ornamentation Development in Music, Principles of its Formation and Performance Style. Вісник Київського Національного Університету Культури І Мистецтв Серія Музичне Мистецтво 3, 169–180. https://doi.org/10.31866/2616-7581.3.2.2020.219167

Child, F.J., Sargent, H.C., Kittredge, G.L., 1904. English and Scottish popular ballads; Boston, Houghton Mifflin co.

Conklin, D., 2006. Melodic analysis with segment classes. Mach. Learn. 65, 349–360. https://doi.org/10.1007/s10994-006-8712-x

Du, J., 2024. Sentiment Analysis and Lyrics Theme Recognition of Music Lyrics Based on Natural Language Processing. J. Electr. Syst. 20, 315–321.

Fitch, W.T., 2015. Four principles of bio-musicology. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140091. https://doi.org/10.1098/rstb.2014.0091

Foscarin, F., Jacquemard, F., Fournier-S’niehotta, R., 2019. A diff procedure for music score files, in: 6th International Conference on Digital Libraries for Musicology. Presented at the DLfM ’19: 6th International Conference on Digital Libraries for Musicology, ACM, The Hague Netherlands, pp. 58–64. https://doi.org/10.1145/3358664.3358671

Haldar, R., Mukhopadhyay, D., n.d. Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach.

Hart, R., 2012. Key-finding algorithm [WWW Document]. URL https://rnhart.net/articles/key-finding/ (accessed 9.24.25).

Herrera, R.J., Garcia-Bertrand, R., 2023. Chapter 6 - Molecular biology tools used to study hominin evolution, in: Herrera, R.J., Garcia-Bertrand, R. (Eds.), Sex and Cohabitation Among Early Humans. Academic Press, pp. 219–279. https://doi.org/10.1016/B978-0-12-819086-9.00012-9

Key | Music, Major, Minor, & Chromatic Scales | Britannica [WWW Document], n.d. URL https://www.britannica.com/art/key-music (accessed 9.24.25).

Madsen, S.T., Widmer, G., n.d. A Complexity-based Approach to Melody Track Identification in MIDI Files.

Mahedero, J.P.G., MartÍnez, Á., Cano, P., Koppenberger, M., Gouyon, F., 2005. Natural language processing of lyrics, in: Proceedings of the 13th Annual ACM International Conference on Multimedia. Presented at the MM05: 2005 13th Annual ACM International Conference on Multimedia, ACM, Hilton Singapore, pp. 475–478. https://doi.org/10.1145/1101149.1101255

Melody | Music Appreciation 1 [WWW Document], n.d. URL https://courses.lumenlearning.com/suny-musicappreciationtheory/chapter/melody-an-overview/ (accessed 9.24.25).

Miller, M., 2020. Komplexitätsanalyse der Klaviermusik europäischer Komponisten anhand von MIDI-Files.

Moore, J.R., 1916. The Influence of Transmission on the English Ballads. Mod. Lang. Rev. 11, 385–408. https://doi.org/10.2307/3712769

OnMusic Dictionary - Term [WWW Document], n.d. URL https://dictionary.onmusic.org/terms/3335-stanza (accessed 9.24.25).

Salmon, P., Newmark, J., 1989. Clinical Applications of MIDI Technology. Med. Probl. Perform. Art. 4, 25–31.

Savage, P.E., Passmore, S., Chiba, G., Currie, T.E., Suzuki, H., Atkinson, Q.D., 2022. Sequence alignment of folk song melodies reveals cross-cultural regularities of musical evolution. Curr. Biol. 32, 1395-1402.e8. https://doi.org/10.1016/j.cub.2022.01.039

Schellenberg, E.G., von Scheve, C., 2012. Emotional cues in American popular music: Five decades of the Top 40. Psychol. Aesthet. Creat. Arts 6, 196–203. https://doi.org/10.1037/a0028024

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Wolfe, J., n.d. Note names, MIDI numbers and frequencies [WWW Document]. UNSW Music Acoust. URL https://www.phys.unsw.edu.au/jw/notes.html (accessed 5.13.25).
Wybren de Jong, J., Shlien, S., 2021. Guide to EasyABC [WWW Document]. sourceforge. URL https://easyabc.sourceforge.net/ (accessed 8.7.24).

Zimmerman, Ri.L., 1997. Musical Notation in the Teaching of Rhythm & Stress. 夙川学院短期大学研究紀要 21, 49–61. https://doi.org/10.24668/shukulib.21.0_49