Relating Natural Language Text to Musical Passages

Richard Sutcliffe
School of CSEE
University of Essex
Colchester, UK
rsutcl@essex.ac.uk

Tim Crawford
Dept of Computing
Goldsmiths, University of London
t.crawford@gold.ac.uk

Chris Fox
School of CSEE
University of Essex
Colchester, UK
foxcj@essex.ac.uk

Deane L. Root
Department of Music
University of Pittsburgh
Pittsburgh, PA, USA
dlr@pitt.edu

Eduard Hovy
Lang Technologies Inst
Carnegie-Mellon Univ
Pittsburgh, PA, USA
hovy@cmu.edu

Richard Lewis
Department of Computing
Goldsmiths, University of London
richard.lewis@ gold.ac.uk

ABSTRACT

There is a vast body of musicological literature containing detailed analyses of musical works. These texts make frequent references to musical passages in scores by means of natural language phrases. Our long-term aim is to investigate whether these phrases can be linked automatically to the musical passages to which they refer. As a first step, we have organised for two years running a shared evaluation in which participants must develop software to identify passages in a MusicXML score based on a short noun phrase in English. In this paper, we present the rationale for this work, discuss the kind of references to musical passages which can occur in actual scholarly texts, describe the first two years of the evaluation and finally appraise the results to establish what progress we have made.

1. INTRODUCTION

A traditional Information Retrieval (IR) system takes as input a short textual query and a document collection and returns a list of documents which match the query [27]. By combining IR with Natural Language Processing (NLP) the field of Question Answering was born [13], leading to systems which could take a query as input and produce an exact answer [17-20,24]. In the meantime, Music Information Retrieval (MIR) has become a very active area in which various kinds of query are matched against music recordings or electronic forms of score such as MEI [11] (inspired by TEI [25]) or MusicXML [15].

However, music involves text as well as scores; there is a vast body of textual information concerned with Western classical music. First and foremost, Grove's Dictionary of Music and Musicians has developed from a four-volume printed dictionary published in 1879-1889 into Grove Online which contains around 50,000 signed articles and 30,000 biographies contributed by over 6,000 scholars [6]. In addition, there are countless scholarly books, journal articles and conference papers as well as numerous online sources such as the Wikipedia. All these sources contain detailed analyses of musical works which necessarily make reference to specific passages in scores. Our long-term objective is to investigate whether these references – expressed in a natural language such as English – can be automatically matched to the musical passages to which they refer.

In pursuit of our objective we organised in 2014 [23, 10] and 2015 [to appear] shared evaluations called C@merata (Cl@ssical Music Extraction of Relevant Aspects by Text Analysis) – http://csee.essex.ac.uk/camerata/ – in which a number of participants each built a system which could take as input a question in English and a score in MusicXML and identify one or more passages in the score which matched the question. We describe those evaluations and the rationale behind them. We first outline the background to this work and its origins in Question Answering (QA). Second, we present an analysis of text examples, taken from the writings of three important musicologists, which refer to musical passages. Third and Fourth we describe the two C@merata campaigns. Finally we discuss what we have learned and draw some conclusions.

2. BACKGROUND TO OUR EVALUATIONS

Our work is derived from three existing areas of research. First, the considerable body of MIR work concerned with finding passages in music scores based on inputs of various kinds, e.g. [5].

Secondly, the Music Information Retrieval Evaluation Exchange has been organised by J. Stephen Downie since 2005 [4,12]. These landmark evaluations have been concerned with many different tasks over the years and are related to parallel evaluations concerning IR and NLP at TREC [26], CLEF [1] and NTCIR [16]. While MIREX has often been concerned with audio-based systems, it has regularly featured score-based tasks which, in the light of our work, could be combined with natural

language input.

Thirdly, there have been QA tracks at CLEF, starting in 2003 [20]. However, these were not concerned with music until 2011. In that year, the Question Answering for Machine Reading (QA4MRE) task featured difficult multiple-choice questions in four domains, one being Music and Society [18]. Four documents in this domain were used, each taken from transcripts of talks delivered at the TED Conferences. In the 2012 task [19], the music texts used were drawn from Wikipedia, Project Gutenburg and the 1911 Encyclopedia Britannica. Finally, in 2013, the four documents were taken with permission from Grove Online [6]. This gave us the idea of combining text processing with core processing.

3. REFERENCES TO SCORES IN MUSIC TEXTS

In this section, we motivate our work by providing a short description of the references to musical passages in three important text sources. The first is an analysis of the Beethoven Symphonies by Antony Hopkins (Chapter 2: *Symphony No. 1 in C Major Op. 21*) [7] (henceforth ah). The second is the study of Domenico Scarlatti by Ralph Kirkpatrick (Chapter 10: *Scarlatti's Harmony*, Section *Cadential vs. Diatonic Movement of Harmony*) [9] (henceforth rk). The third is the entry for Anton Bruckner by Deryck Cooke (*Section 7. Music*) [2] from the New Grove Dictionary of Music and Musicians [21] (henceforth dc).

We extracted phrases from the above works by hand – 261 in all – and organised them into 14 categories: notes, intervals, scales, melodies, rhythms, tempi, dynamics, keys, harmony, counterpoint, texture & instrumentation, bar numbers, passages & sections and structures & sequences. Furthermore, they are classed as Specific or Vague. Examples of each category can be seen in Table 1, with two Specific and two Vague for each phrase type. The source is indicated in square brackets: [ah26] means ah (i.e. Hopkins) p26; [dc364lh] means dc (i.e. Cooke) p364 in Grove, left hand column.

It is important to note that the categories in Table 1 are for illustration only and are neither exhaustive nor mutually exclusive. The examples are given purely to illustrate the kinds of references to musical passages which one might find in a musicological text. Moreover, the binary categorisation into Specific and Vague is also purely for illustration purposes as specificity lies on a scale. We now draw some conclusions from this table.

The first point to note is that the references vary in specificity; some are clear and unambiguous (C#-D rising semitone, D major, eight-part choir, bars 189-198); others are much more difficult to pin down (alien F#, disturbing syncopations, anguished D minor chromaticism, varied alternation of two long-drawn themes). Secondly, however, all the phrases are meaningful – an expert familar with the works concerned is likely to be able to identify the points mentioned in the score with a fair accuracy (high Precision even if not necessarily high Recall). This suggests that they are interesting and worthwhile to study.

Thirdly, some categories of phrase lend themselves to rather simple and clear expression. Examples include Notes (G), Intervals (ascending diminished fifth), Scales (D major), Rhythms (repeated crotchet chords), Dynamics (FF), Keys (Bb minor) and bar numbers (measure thirteen). If we set ourselves the task of

Category	S/V	Examples
Notes	S	[ah26] giant unison G from the entire orchestra
		[rk220] based on nothing else but A, D, E, and A
	V	[ah12] alien F# in the ascending scale
		[dc364lh] pedal point
Intervals	S	[ah24] C#-D rising semitone
		[dc363lh] an ascending diminished fifth
	V	[ah19] fragment of five rising crotchets
		[dc364lh] themes based on falling octaves
Scales	S	[dc363lh] parts entering successively on the degrees of the ascending scale of D major
		[dc363rh] old church modes ... Phrygian and Lydian
	V	[ah28] the initial scale
		[ah29] little scales dart to and fro
Melodies	S	[ah13] semiquaver descent in bar 18
		[ah19] fragment of five rising crotchets
	V	[ah19] Second Subject appearing in the tonic key
		[dc363rh] the chorale themes in the symphonies
Rhythms	S	[ah15] quaver pattern
		[ah25] repeated crotchet chords
	V	[ah18] disturbing syncopations
		[dc364lh] hammering ostinatos
Tempi	S	[ah11] slow tempo
		[dc366lh] slow movements
	V	[ah28] rustic oom-pah bass
		[dc364rh] intense and long-drawn string cantabile
Dynamics	S	[ah26] violins in bar 126 come in FF
		[ah29] sudden fortissimo outburst
	V	[ah29] sudden roaring
		[dc364lh] murmering tremolando
Keys	S	[ah10] D major
		[dc366lh] in Bb minor
	V	[rk221] modulatory excursion of the second half
		[dc363rh] unusual key changes
Harmony	S	[rk221] major dominant
		[dc364lh] tonic triad of E major
	V	[rk220] departure from three-chord harmony
		[dc363lh] anguished D minor chromaticism
Counterpoint	S	[ah23] cellos provide a delicate countertune
		[dc363lh] parts entering successively on the degrees of the ascending scale of D major
	V	[rk220] dominated by diatonic movement of parts
		[dc363lh] bold polyphonic imitation of a single point
Texture, Instrumenation	S	[dc363lh] eight-part choir
		[dc363rh] a piece of unison plainsong
	V	[ah29] decked with garlands of scales from flutes, clarinets and bassoons
		[dc364rh] a faint background sound, emerging almost imperceptibly out of silence
Bar Numbers	S	[ah15] bars 189-198
		[rk220] measure thirteen to measure fifteen
	V	[ah24] sixteen or at most thirty-two bars long
		[dc365rh] over periods of 16, 32 or even 64 bars
Passages, Sections	S	[dc363rh] whose slow movement and finale
		[dc364rh] far-ranging first movement
	V	[rk220] series of small sequential passages
		[dc362rh] a passage from the Gloria
Structures, Sequences	S	[ah18] First Subject
		[rk221] Phrygian cadence
	V	[dc365rh] exposition (nearly always built on three subject groups rather than two)
		[dc366rh] varied alternation of two long-drawn themes

Table 1. Fourteen types of referring expressions, categorised into Specific (S) and Vague (V).

searching for such passages in a score, we are likely to be quite successful.

Fourthly, some categories of phrase tend conversely to be complex and often imprecise as well. Examples include Texture & Instrumentation (a faint background sound, emerging almost imperceptibly out of silence), Passages & Sections (a passage from the Gloria) and Structures & Sequences (exposition (nearly always built on three subject groups rather than two)). Western classical music excels in structure and in harmony, so treatment of these topics tends to be particularly interesting and important. The richness and ambiguity of language are its strengths in this context as a great deal can be suggested in relatively few words. Moreover, to the expert, the references remain quite clear, though a considerable amount of knowledge and background information is being brought to bear.

Fifthly, it is interesting to observe that many of the examples in Table 1 are noun phrases; this construct can express very complicated and detailed concepts in a musicological text.

Sixthly and finally, phrases in natural language can never be replaced by expressions in a pattern language (such as regular expressions applied over text strings). Such expressions are by their nature unambiguous and in practical contexts they are usually concise. Therefore, the study of natural language in musicology is not made unnecessary by the existence of such languages. On the other hand, such expression languages are extremely useful and worthwhile [28]; one possible application of them here is to map a natural language phrase onto a pattern (possibly extremely complex) in such an expression language in order to initiate a search.

In the next section we will describe our evaluations.

4. THE 2014 C@MERATA TASK

4.1 Input Provided

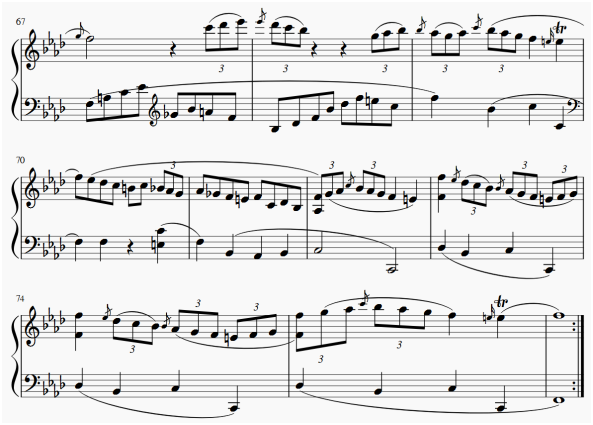
In a QA evaluation such as ResPubliQA [17], the input is normally a short question such as 'Who is President of the United States' and the output is an exact answer such as 'Barack H. Obama'. As we have discussed earlier, many of the real examples in Table 1 are in fact noun phrases. So it seemed reasonable to use a noun phrase as the input for an initial evaluation, rather than a complete question. The top of Table 2 shows the question types which were adopted. For all the types mentioned below, there are several examples in the right hand column.

As we observed above, entries in the Notes category of Table 1 are some of the simplest and clearest. As this was a new task, it was decided to include three simple query types in the evaluation which correspond broadly to Note: simple_pitch, simple_length and pitch_and_length.

Perf_spec queries combine a note with some performance indication. Stave_spec queries restrict the answer to a particular stave in the score which may be specified in various ways, including the instrument concerned, the hand being used (for keyboard music) or the clef on which the music appears. Similarly, word_spec queries link a note to the word which is sung on it in one of the parts.

Question Types for 2014 Task		
Type	**No**	**Examples**
simple_pitch	30	G5, E, A natural, C flat, F#4, F2 sharp
simple_length	30	dotted quarter note, quarter note rest, semiquaver rest, whole note, semibreve
pitch_and_length	30	D# crotchet, half note C, quarter note B5, semiquaver G#, half note Db, quaver F#
perf_spec	10	D sharp trill, fermata A natural, staccato B flat, marcato D flat, F trill, down bow E
stave_spec	20	D4 in the right hand, half note D in the viola, treble clef A sharp, F3 sharp in the ″alt″, quarter note F in the Alto
word_spec	5	word "Se" on an A flat, minim on the word ″Der″, minim B on the word ″im″, G on the word ″praise″
followed_by	30	crotchet followed by semibreve, D followed by G, quarter note G followed by eighth note G, dotted quaver E followed by semiquaver F sharp, crotchet rest followed by crotchet, dotted quarter note followed by A4
melodic_interval	19	melodic octave, rising major sixth, melodic descending fifth, falling major third, melodic rising minor third, octave leap, falling tone, melodic fourth
harmonic_interval	11	harmonic major sixth, harmonic second, nineteenth, seventh, harmonic fifth, harmonic octave, major seventeenth
cadence_spec	5	perfect cadence
triad_spec	5	tonic triad, Ib triad, triad in first inversion, Ia triad
texture_spec	5	polyphony, melody with accompaniment, monophony, homophony
All	**200**	
Question Types for 2015 Task		
Type	**No**	**Examples**
1_melod	40	D4 minim, eighth note in measure 9
1_melod qualified by perf, instr, clef, time, key	40	trill on a quaver A; G# in the Cello part in measures 29-39; sixteenth note C# in the left hand; half note E3 in 2/2; sixteenth note G in G minor in measures 1-5
n_melod	20	F# E G F# A; Do Mi Do Sol Do Mi Sol Do in bars 1-20; twenty semiquavers; five note melody in bars 1-10
n_melod qualified by perf, instr, clef, time, key	20	two staccato quarter notes in the Violin 1; crotchet, crotchet rest, crotchet rest, crotchet, crotchet rest, crotchet, crotchet, crotchet, crotchet, crotchet in the Timpani; melodic octave leap in the bass clef in measures 70-80; G4 B4 E5 in 3/4; rising G minor arpeggio
1_harm possibly qualified by perf, instr, clef, time, key	20	eighth note chord Bb, C, E; chord of D minor in measures 109-110; harmonic minor sixth in the Violas; dotted minim chord in the left hand
texture	6	monophonic passage; homophony in measures 1-14; polyphony in measures 10-14; Alberti bass in measures 0-4
follow possibly qualified on either or both sides by perf, instr, clef, time, key	40	quavers F4 E4 in the oboe followed by quavers E2 G#2 in the bass clef; quarter note minor third followed by eighth note unison; C followed by mordent Bb; chord C4 G4 C5 E5 then a quaver; three eighth notes in the Violin I followed by twelve sixteenth notes in the Violin II in measures 87-92
synch possibly qualified in either or both parts by perf, instr, clef, time, key	14	four eighth notes against a half note; crotchet D3 on the word "je" against a minim D2; four staccato quavers in the Violoncello against a minim chord Ab3 C4 F4 in the Harpsichord
All	**200**	

Table 2. Summary of question types in tasks.



Q: G flat
A: [4/4, 2, 67:5-67:5], [4/4, 2, 71:2-71:2]
Q: semibreve
A: [4/4, 1, 76:1-76:4]
Q: minim F
A: [4/4, 1, 67:1-67:2]
Q: minim C in the bass
A: [4/4, 1, 72:1-72:2], [4/4, 1, 72:3-72:4]
Q: crotchet followed by semibreve
A: [4/4, 1, 75:4-76:4]
Q: melodic octave
A: [4/4, 2, 69:5-69:8], [4/4, 2, 72:1-72:8], [4/4, 2, 73:5-73:8], [4/4, 2, 74:5-74:8], [4/4, 2, 75:5-75:8]

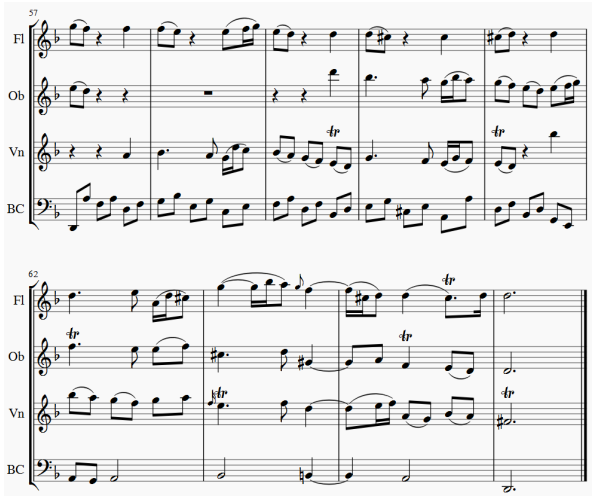
Figure 1. Extract from Scarlatti K466 with questions and answers from the 2014 task.

So far, all the query types are simple notes in isolation. Queries of type followed_by specify two adjacent notes.

As Table 1 showed, intervals are discussed in real texts, so we wished to include some queries of this type. We divided them into two kinds, melodic and harmonic. melodic_interval specifies two adjacent notes on the same stave which are a specified distance apart. Conversely, a harmonic_interval specifies two simultaneous notes. Unlike melodic intervals, harmonic intervals were permitted to occur across staves because they are integral to the concept of harmony which is often created by instruments or voices in different parts. Intervals are considered harmonic by default, thus 'fifth' is assumed to be a harmonic fifth.

The last three question types were more experimental, though still being relatively straightforward and unambiguous in musical terms. cadence_spec requires a cadence to be identified. A triad_spec specifies triads in various forms of notation. Finally, texture_spec states the required texture to be found. Referring back to Table 1, cadences touch upon Structures & Sequences and Triads are a fundamental element of Harmony.

There were 200 queries in a fixed distribution as shown in the middle column of Table 2. The four simplest query types (simple_pitch, simple_length, pitch_and_length, followed_by) were the most numerous in the test set with 30 each. After this came stave_spec and melodic interval with twenty each followed by perf_spec and harmonic_interval with ten each. (One melodic interval was changed for a harmonic_interval at a late stage, so in fact there were nineteen of the former



Q: dotted minim F#4
A: [3/4, 1, 65:1-65:3]
Q: F4 crotchet in the oboe
A: [3/4, 2, 64:3-64:4]
Q: minim A2 in 3/4 time
A: [3/4, 1, 62:2-62:3], [3/4, 1, 64:2-64:3]
Q: chord D2 E5 G5 in bars 54-58
A: [3/4, 2, 57:1-57:1]
Q: quavers F3 A3 followed by crotchet A4 in the violin
A: [3/4, 1, 57:2-57:3]
Q: four quavers in the violin against a minim in the bass clef
A: [3/4, 1, 62:2-62:3], [3/4, 1, 64:2-64:3]

Figure 2. Extract from Bach BWV1047 Andante with questions and answers from the 2015 task.

and eleven of the latter.) Finally, there were five each of word_spec, cadence_spec, triad_spec and texture_spec. Thus some more experimental types of query were represented in the task but played a relatively minor role.

In summary, most of the question types used in 2014 were straightforward and were derived from Notes, Intervals and (partly) Harmony, Texture & Instrumentation and Structures & Sequences. Other phrase types of Table 1 were not catered for.

4.2 Output Required

As we have seen, an input query was simply a short noun phrase. To make the evaluation as simple as possible, an answer was defined to be a subsection of a score, starting and ending at a particular place. The answer was not required to specify which stave (or staves) contained the answer.

Initially, we planned to measure beats in a bar in terms of the shortest note (hemidemisemiquaver, one sixteenth of a crotchet). However, this does not allow for triplets (where, say, a crotchet is divided into three) or any other sort of n-tuplet. So instead, we adopted the concept of *divisions* from MusicXML. The divisions value is the number of beats into which the crotchet is divided. A suitable value depends on what we wish to demarcate as an answer. So for simplicity, we specified for each query the divisions value to be used for the answers.

Based on these ideas we developed the concept of a *passage* which would contain, for both start and end, a time signature, a divisions value, and a bar and beat.

The start bar and beat is where the passage is defined to commence. More precisely, the passage begins in the denoted bar immediately *before* the start beat, measured from the beginning of the bar in the unit of time denoted by the stated divisions value. Similarly, the passage is defined to end immediately *after* the end beat. We adopted this before-the-start and after-the-end after careful thought and discussion. The advantage of it is that it is intuitive: As can be seen in Figure 1, above, the first two crotchets in bar 67 are denoted 67:1-67:2 which can be understood at a glance.

We developed three equivalent ways of stating a passage: Ascii Long Form, Ascii Short Form and XML form. The Ascii forms are convenient for discussions in papers etc. while the XML form is useful as the input to, and output from programs.

Here is an example in short form: [4/4,1,1:1-2:4]. The time signature is 4/4 and divisions value is 1. The passage starts in bar 1 before the first crotchet (i.e. 1:1) and ends in bar two after the fourth crotchet (i.e. 2:4). We take bar numbers from the MusicXML score.

We use the XML format for specifying the test queries for participants as well as for the queries plus correct answers (often called the Gold Standard in QA).

In summary, our passage specifies two vertical lines drawn through the score and does not distinguish between the different staves. We thus assume that any answer can be exactly demarcated in this way. We will return to this point in the conclusions.

4.3 Evaluation

Precision, Recall and F-Measure are commonly used in IR and NLP [27]. We wished to determine all the correct answer passages by hand to produce a Gold Standard and then to compare the results returned by a system to that.

It is useful to have both strict and lenient measures in an evaluation. At the fourth TREC QA track onwards (starting in 2002) there were four judgements of each answer, Right, ineXact, Unsupported and Wrong [29]. In the TREC context a correct answer could be 'Bill Clinton' while an ineXact one could be 'Clinton' or perhaps 'Bill Clinto'. Unsupported answers were Right but not shown to be so from a document in the collection.

We decided that a passage returned which began at the right bar and beat within the bar and also ended at the right bar and beat within the bar was correct. On the other hand, an answer which started and ended at the right bar (but not necessarily the right beat in the bar) was still very useful and could be considered the equivalent of TREC's ineXact. If an expert is looking for a particular cadence, for example, and is told the bar numbers, they can usually see it at a glance. However, searching through hundreds of bars looking for the cadence is time consuming. The concept of Unsupported is not applicable to our task. The measures were thus defined as follows:

Beat Precision (BP) is the number of beat-correct passages returned by a system divided by the number of passages (correct or incorrect) returned.

Beat Recall (BR) is the number of beat-correct passages returned by a system divided by the total number of answer passages known to exist.

2014 Results	BP	BR	BF	MP	MR	MF
Maximum	0.713	0.904	0.797	0.764	0.967	0.854
Minimum	0.113	0.150	0.185	0.155	0.154	0.226
Average	0.420	0.654	0.483	0.460	0.734	0.534
2015 Results	**BP**	**BR**	**BF**	**MP**	**MR**	**MF**
Maximum	0.817	0.739	0.620	0.817	0.809	0.656
Minimum	0.061	0.175	0.108	0.073	0.175	0.129
Average	0.351	0.564	0.348	0.370	0.619	0.375

Table 3. Results of the 2014 & 2015 tasks.

As is usual, **Beat F-Score** (BF) is the harmonic mean of BP and BR.

Measure Precision (MP) is the number of bar-correct passages returned by a system divided by the number of passages (correct or incorrect) returned.

Measure Recall (MR) is the number of bar-correct passages returned by a system divided by the total number of answer passages known to exist.

Finally, **Measure F-Score** (MF) is the harmonic mean of MP and MR.

4.4 Scores

After consideration of several notations including kern [8], MusicXML was chosen because it is widely used and is supported by music21 [3] and musescore [14].

Twenty MusicXML scores were used and ten questions were set on each, forming the question type distribution of Table 2. We incorporated both European (crotchet, bar etc) and American (quarter note, measure etc) terms into the task by setting American queries for ten of the twenty scores and English queries for the rest.

Scores for 2014 were chosen from the Renaissance and Baroque in order to avoid more heavily-scored works from the Classical period onwards. The composers chosen were Bach, Carissimi, Charpentier, Corelli, F. Cutting, Dowland, Lully, Monteverdi, Purcell, A. Scarlatti, D. Scarlatti, Tallis, Telemann, Vivaldi and S. L. Weiss. Scores were chosen on a predefined distribution: six on two staves, six on three staves, four on one stave and two each on four staves and five staves. There were works for solo cello, harpsichord and lute; one, three, four and five voices; soprano or cello and harpsichord; two violins and cello; two violins, viola and two cellos.

The scores were obtained from two sources. Most came from musescore.com. Two Bach chorales were used and both came from www.jsbchorales.net. We required scores to have a license 'to share' rather than just 'for personal use'. Moreover, we required scores to be well presented, transcribed in a scholarly manner and provided in valid MusicXML Version 2 or lower.

4.5 Questions

Each score was sent to one of the organisers who was asked to set questions according to the target distribution of Table 2. It was specified for each score whether the questions were to be in American or English. For each question, answers were to be provided in the Ascii short form for specifying passages. The organiser in question

was asked to find all answers for all the questions. The question data was returned in an Ascii format which incorporates the score filename, the questions, the answers in Short Ascii form and also any comments concerning the questions or answers.

On receipt of the files, the questions and answers were checked by a second expert who noted any changes or observations using comments in the Ascii file. The second expert also carried out an independent search for answer passages within the scores. When all changes were checked and validated, the complete set of twenty Ascii files was transformed automatically into XML format in order to form the Gold Standard for the task.

4.6 Participants, Runs and Results

The task was announced in January 2014. Five participants registered; two were from Ireland and the other three came from Australia, England and India. Participants had one week to complete their runs starting from 16th June 2014.

Each participant was allowed to submit up to three runs. The overall results are shown in Table 3. The best BF (strict) score was 0.797 which was remarkably good.

Averages for BF and MF are 0.483 and 0.534 so systems scored better under lenient measures than under strict measures but the difference is not large – only 11%. Concerning the top run, the difference between MF=0.854 and BF=0.797 is only 7%. So if a system finds the correct bar, it tends to find the exact beat in the bar as well. Generally, the average figures suggest that participants had all made a very good attempt at building a system for this very complicated task.

4.7 Approaches to the Task

Concerning software, most participants opted to use Python and to adapt a baseline system using music21 [3] which we wrote and distributed [22]. Others used their own tools in Lisp or C.

Only basic NLP was used. Typically the query was scanned looking for terms (e.g. down bow) and converting them to concepts (down_bow). Some systems adopted a QA approach and assigned the query to a pre-define set of types, each with its method of solution. Others converted the concepts to a structured representation by parsing the concepts. The final stage was a search of the score. Some varied the representation of the score according to the query type (e.g. using music21 chordify for cadence questions). As all answers to a given query were defined to lie in exactly one of the scores, no one opted to use any inverted indexing of the music data.

5. THE 2015 C@MERATA TASK

5.1 Changes from 2014

This year's campaign has just concluded. The use of MusicXML scores, the XML formats for questions and answers, the passage concept and the evaluation measures remained the same in 2015. However, there was a wider range of score types from the Renaissance to the early Romantic periods, scores were more complicated – up to nineteen staves – and questions were differently organised and generally more difficult (see Table 2). For example, an n_melod question can specify quite complicated melodies while the synch type can link two simultaneous features.

5.2 Participants, Runs and Results

The same five participated as in 2014. The maximum BF was 0.620 and the average BF was 0.348 (Table 3), both lower than last year. However, the task was considerably harder and the participants did very well.

6. DISCUSSION AND CONCLUSIONS

First, in both years, participants were able to build a working system and submit valid runs.

Second, all systems could make a good attempt at answering at least one of the question types.

Third, the best systems (see Table 3) achieved very good results and several others were not far behind.

Fourth, the technical basis of the task was shown to be sound and all the steps of the campaigns were fulfilled.

Fifth, the development of strict measures (BP, BR, BF) and lenient measures (MP, MR, MF) specifically for this task worked well.

Sixth, the ability to evaluate runs automatically showed the practicality and scalability of the evaluation.

There were also some shortcomings; first, our passage concept does not distinguish between staves. Suppose a minim F starts in the first beat of bar 1 in the treble clef and in the second beat of bar 1 in the bass clef (of a keyboard work). The two answer passages thus overlap which is anomalous. On the other hand, consider a texture such as homophony where some instruments have rests for some or all of the passage – are those instruments part of the passage or not?

Second, not all passages of interest in a score can be demarcated exactly. For example, a polyphonic passage may commence in a madrigal when a homophonic section is still drawing to a close. If we say 'most' parts must be participating in polyphony is that the start of it, or must 'all' participate? Also, what about the start and end of a triad? Sometimes the bass note is only established after the other notes.

Third, some 'passages' may turn out to have no length. Consider the perfect cadence. The V chord can be set up in many different ways such that it can be hard to say where exactly that chord starts. Then, the onset of the I chord can be equally ambiguous: there may be a trill on the V; or the I in the treble may be set up either before or after the bass moves to I. For the future, we would consider defining a perfect cadence as a point in a score, not a passage; the instant the bass moves from V to I.

Finally, consider again Table 1 (real phrases) against Table 2 (actual phrases used in 2014 and 2015). There is a considerable difference in complexity and subtlety. Many of our queries were simple notes which present few problems for either NLP or MIR. Future campaigns can include more complex query types which delve further into the subtleties of musical language while still being practical for use in MIR.

7. REFERENCES

[1] CLEF (2014). http://www.clef-initiative.eu/.

[2] Cooke, D. (1995). Bruckner, (Joseph) Anton. In S. Sadie (ed), New Grove Dictionary of Music and Musicians, Volume 3, Section 7. Music (p362-366). London, UK: Macmillan.

[3] Cuthbert, M. S., & Ariza C. (2010). music21: a toolkit for computer-aided musicology and symbolic music data. Proc. International Symposium on Music Information Retrieval (Utrecht, The Netherlands, August 09 - 13, 2010), p637-642.

[4] Downie, J. S. (2008). The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research.Acoustical Science and Technology 29 (4): 247-255. Available at: http://dx.doi.org/10.1250/ast.29.247

[5] Ganseman, J., Scheunders, P., & D'haes, W. (2008). Using XQuery on MusicXML databases for musicological analysis. Proc. International Symposium on Music Information Retrieval, p433-438.

[6] Grove Music Online (2015). http://www.oxfordmusiconline.com/public/

[7] Hopkins, A. (1982). The Nine Symphonies of Beethoven. London: Pan Books.

[8] Huron, D. (1997). Humdrum and Kern: Selective Feature Encoding. In 'Beyond MIDI', ed. E. Selfridge-Field (p375-401). Cambridge, MA: MIT Press.

[9] Kirkpatrick, R. (1953). Domenico Scarlatti. Princeton, NJ: Princeton University Press.

[10] Larson, M., Riegler, M. A., Miro, X. A., Korshunov, P., Petkos, G., Soleymani, M., Choi, J., Schedl, M., Ionescu, B., Eskevich, M., Jones, G., & Sutcliffe, R. F. E. (2014). Proc. MediaEval 2014 Workshop, Barcelona, Spain, October 16-17 2014. http://ceur-ws.org/Vol-1263/.

[11] MEI (2014). Music Encoding Initiative. http://music-encoding.org/home.

[12] Mirex (2014). http://www.music-ir.org/mirex/wiki/ MIREX_HOME

[13] Mollá, D., & Vicedo, J. L. (2007). Question answering in restricted domains: An overview. Comput. Linguist., 33(1):41-61.

[14] Musescore (2014). Music Composition and Notation Software. http://musescore.org/.

[15] MusicXML (2014). http://www.musicxml.com/.

[16] NTCIR (2014). http://research.nii.ac.jp/ntcir/index-en.html.

[17] Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N., & Osenova, P. (2009). Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation Notebook of the Cross Language Evaluation Forum, CLEF 2009, Corfu, Greece, 30 September - 2 October.

[18] Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Sutcliffe, R., Forascu, C., Sporleder, C. (2011). Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. Proc. QA4MRE-2011. Held as part of CLEF 2011.

[19] Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P. (2012). Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. Proc. QA4MRE-2012. Held as part of CLEF 2012.

[20] Peñas, A., Magnini, B., Forner, P., Sutcliffe, R., Rodrigo, A., & Giampiccolo, D. (2012). Question Answering at the Cross-Language Evaluation Forum 2003-2010. Language Resources and Evaluation Journal, 46(2), 177-217.

[21] Sadie, S. (eds) (1995). The New Grove Dictionary of Music and Musicians. London, UK: Macmillan.

[22] Sutcliffe, R. F. E. (2014). A Description of the C@merata Baseline System in Python 2.7 for Answering Natural Language Queries on MusicXML Scores. University of Essex Technical Report, 21st May, 2014.

[23] Sutcliffe, R. F. E., Crawford, T., Fox, C., Root, D. L., & Hovy, E. (2014). The C@merata Task at MediaEval 2014: Natural language queries on classical music scores. Proc. MediaEval 2014 Workshop, Barcelona, Spain, October 16-17 2014. http://ceur-ws.org/Vol-1263/.

[24] Sutcliffe, R., Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Forascu, C., Benajiba, Y., & Osenova, P. (2013). Overview of QA4MRE Main Task at CLEF 2013. Proc. QA4MRE-2013.

[25] TEI (2014). Text Encoding Initiative. http://www.tei-c.org/index.xml.

[26] TREC (2014). http://trec.nist.gov/.

[27] van Rijsbergen, K. J. (1979). Information Retrieval. London, UK: Butterworth. http://www. dcs.gla.ac.uk/Keith/ Preface.html

[28] Viglianti, R. (2015). Enhancing Music Notation Addressability. http://mith.umd.edu/ research/ project/ enhancing-music-notation-addressability/.

[29] Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. http://trec.nist.gov/pubs/trec11/ papers/QA11.pdf