

The collection and use of a descriptive corpus for the study of musical effect

Dave Billinge, Department of Creative Technology, University of Portsmouth,
Buckingham Building, Burnaby Road, Portsmouth PO1 3AE.
Email dave.billinge@port.ac.uk

Abstract

This paper describes the structure and methods of a series of experiments carried out to study the use of figurative language in the description of musical effect. As such it focuses on practical as much as theoretical issues, these being discussed extensively elsewhere. It was first necessary to collect a set of descriptive words, refine the set by usage frequency and then analyse responses to the further use of this refined set. As a direct result of these exercises it was decided to extend the study to cover consideration of word groups. The author critically reviews the methodological processes chosen. For a review of experimental outcomes and further theoretical discussion see Billinge (2001) and Billinge and Addis (2001, 2002a, 2002b, 2003).

1. The Aim of the Experiments

Given the extent of shared informal talk and informal writing about the experience of musical performance, it was felt worthwhile to attempt a confirmation that listeners successfully communicate their feelings and to clarify which mode of linguistic communication they use. If successful the results could contribute to the creation of expert systems in artistic decision-making. Section 6 discusses briefly how theory has been revised in this respect.

2. Collecting the vocabulary

2.1 The Lexicon

The purpose of the first experiment was to identify a lexicon of descriptive words used by music lovers. To stage any experiments on the descriptions used by listeners there had to be a preliminary set of words. Lacking any previously established lexical set for such use, the choice lay between creating a list of one's own, taking it from the most easily available published source, or collecting it from the users. The creation of a list oneself was dismissed as too subject to bias. Given that any initial request for words would have to make allowance for the unwillingness of respondents in the mature age group targeted to commit themselves without guidance, it was decided that some examples had to be included. The author listed a set of words from several randomly sampled copies of *Gramophone* magazine¹. *Gramophone* is the closest this collecting field cum hobby has to a trade paper, and revised it with the help of a colleague so as to remove at least some of the personal bias. The questionnaire included a list of 50 musical compositions. Knowing the sensitivity of the music lover to assessment of his or her knowledge of the orchestral repertoire this list was not a simple matter. It could not be all popular music because of the bias that would place on the type of music listed. The popular classical repertoire is largely late Classical and Romantic music, and promoters tend to avoid contentious or "difficult" music because concert promotion is a commercial act. The chosen list therefore had to stray a little away from this central repertoire without making any respondent feel ignorant by asking for reactions to pieces of music of which they had never heard.

The instructions included the following sentences.

If, for any one piece, you cannot think of any words, then please refer to the list on the back page for inspiration, but I am much more interested in your own words. It is quite possible that your words are in my list already; this does not matter in the slightest! Finally, it does not matter if you use the same words several times, the order in which you enter words does not matter and "Word 1, 2, and 3" are only there to guide you.

¹ Gramophone magazine is currently celebrating its 80th year of publication. Since 1923 it has published a monthly review of primarily classical music recordings. It is thus seen as the most important international publication of its kind.

Table 1 shows a sample of the chosen 50 musical items in the layout actually issued.

		Word 1	Word 2	Word 3
22	Ibert: Divertissement			
23	Mahler: Symphony No.2 "Resurrection"			
24	Mendelssohn: Violin Concerto			
25	Mendelssohn: A Midsummer Night's Dream			
26	Mozart: Symphony No.40 in G minor			
27	Mozart: Eine Kleine Nachtmusik			
28	Mussorgsky/Ravel: Pictures at an Exhibition			
29	Nielsen: Sinfonia Espansiva			
30	Prokofiev: Lieutenant Kijé			

Table 1: First Questionnaire (extract)

The list of words from *Gramophone* was appended so that no one need feel unable to give a response. This was also for the purposes of keeping the respondents cooperative because many were needed for subsequent work.

impulsive	labyrinthine
individual	lacklustre
inspired	light
intricate	lightweight
inventive	lively
involved	long breathed
inward	loving
joyous	luminous
keen	lurid
kitschy	lusty

Table 2: Given vocabulary (sample)

It was hoped that this approach would result in an experimental corpus that had high user acceptability. Word frequencies were used to reduce the resulting set of 1032 words to a manageable size. Words appearing less than six times were not utilised because in common with all word usage distributions (Zipf 1949) the numbers of words repeated just a few times are huge. It is as the rate of repetition rises that the items appearing with such frequencies grow smaller. Six was chosen as the cut off point because the number of words repeated three times (33), four times (22) and five times (13) were much larger than the number repeated six times (only 8) and would thus have made membership of the "common" set skew disproportionately to less frequent words.

2.2 The Respondents and the Responses

The 12 respondents volunteered from a group of about 60 attending a music day school. This initial group was smaller than intended because of administrative problems with the organisers who, oddly, considered the author's questionnaire an attempt to use their customers without their permission. This is mentioned here in a methodological discussion as a warning to those focussed on what they see as an innocent academic pursuit that not all those involved necessarily see it that way. Later experiments were better prepared in this respect and much higher responses achieved, thus any restrictions inadvertently applied to the initial vocabulary set was overcome subsequently.

No attempt was made to gain a balance by ages or sex because of the profile of attendees and the unfortunately restricted size of the group. Those involved here and later throughout the study were

representative of the local concert-going public in that they tended to be middle-aged or elderly rather than young, though youth was not actively excluded. The author bore in mind the possibility that age, education and sex might be significant in an exercise so closely allied to vocabulary size but possibly because the sample was too small, no differences arose in respect of these characteristics.

Finally in respect of personnel it should be noted that the author achieved insightful and instantaneous feedback from one volunteer who said that a request for three discrete words was not nearly enough to allow their feelings to be expressed. “Simplistic if not positively foolish” was the phrase actually used. In retrospect this pinpointed not so much an experimental design flaw as a weakness in theory subsequently acknowledged as this research has moved away from the discrete lexicon to embrace phrasal, figurative language.

To avoid loss of potentially valuable data several non-lexical facts were recorded. It was not known whether vocabulary would vary by sex or age so this was recorded. Though respondents were explicitly instructed to ignore word order (see Table 1 above) this positioning was recorded so that account could be taken if later analysis implied it to be important.

A certain amount of personal judgement and editing was also needed. Some words were used incorrectly. For example Bartók’s *Music for Strings Percussion and Celesta* is not *atonal* but was so described. Despite being factually incorrect the word was admitted as a figurative usage. Word misuse had to be considered, for example it was decided that *emotive* probably meant *emotional*. Such errors were simply corrected. One respondent noted in the margin that by *varied* she meant *serious to romantic* and *amusing, dramatic*. This way of getting in more than the required number of words was accepted and the words added to the tally.

3. The First Group Experiment

3.1 Data Recording

The first group experiment utilized the above vocabulary set to explore the extent of user agreement on musical predication. Nineteen people participated in three groups on different dates including one with markedly younger testees. The questionnaire asked for a few personal details and the results anonymised and summarised as in Table 3.

Sex	Age	Occasional Listener	Regular Listener	Frequent Listener	Instrumental Player	Participant Number
F	49			X	Y	1
M	56		X		N	2
F	44	X			Y	3

Table 3: Participant Data (extract)

Sex was noted, as above, because it was possible there would be differences in vocabulary choice between men and women. Secondly the participant’s age was noted. There was no evidence to support the prediction that older and younger people would choose from different vocabulary sets but it could not be excluded. The third question concerned a grading of listening experience from “Occasional” to “Frequent” listener. The assumption here was that the more experienced listener would be more likely to have read promulgations of this vocabulary in the journals and newspaper sections devoted to it and thus been influenced more. As it turned out the agreements detected were so small that such subtle analyses were redundant at best. Finally it was asked if the participant played an instrument. Since the aim of this research was the investigation of non-technical language it seemed sensible to assume that knowledge of the technical vocabulary would be influential.

Prior to commencing the experimental sessions it was emphasized that participants should not discuss anything with their fellows until specifically asked to do so. The purpose of the session was explained and that the discussions after the fourth test of this session would be recorded. At the end of the tests the composers and titles were revealed because everyone wanted to know “the answers”. To avoid any

future bias no comment was passed by the author, beyond expressions of satisfaction, to indicate what he thought about the discussions he had heard.

This was the first opportunity to record a real linguistic corpus actually focused on the experimental task, communicating feelings about music in the focal language. As such this was expected to be valuable. It has been the authors experience that effort put into the quality of the recording medium is quickly repaid. It is easy to make a bad recording in which potentially vital data is lost through inadequate signal to noise ratios, background disturbance or even recording pitch instability. Hidden microphones are very unlikely to pick up subtle inflexions because they are too remote from the speaker. Experimental participants on these occasions were told that recording would be made and a good quality stereo microphone was hung, studio style, over the meeting table. Levels were checked beforehand on a professional standard cassette machine² (today we should use a digital medium) and the much-abused Dolby noise reduction was correctly applied on high quality tape. This effort paid off because every nuance of some prolonged conversations could now be used and reused without the effort of listening being at all burdensome. A considerable amount of the most valuable figurative language was used in quiet asides between participants, all captured clearly on tape. Whilst not exactly a methodological issue, the author believes investigators of natural language usage overlook the issue of fidelity in audio recording at their peril.

Similarly the music itself was well reproduced on a high quality system. The author reasoned that if emotional responses were sought then it would be better to ensure that there were as few distortions of reality as possible to clear the way for that reaction.

3.2 The Exercises

The sessions were divided into four exercises. Each of these short “tests” were designed to place increasing pressure on the participants to agree, culminating in an explicit demand for agreement. Each test description is followed by a short rationale.

Test 1 required each participant to listen to ten short items of classical music, mostly extracts, and write down a one word descriptive response without discussion. This produced a freely chosen list without the influence of others. Sufficient time was given for all to finish without pressure being applied.

Test 2 consisted of a replay of the same items but this time the participant had to chose one word from a given list. This provided the experimenter with a set of results that, because the range of words was restricted, had an increased likelihood of displaying agreement. Again no discussion was allowed.

Test 3 consisted of ten new sections of music with a given accompanying descriptive word. Participants were asked to say whether they agreed / disagreed on a five-point scale with the appropriateness of the given word. This test enforced even more restrictions in that just a single word was available and only its appropriateness had to be decided. Agreement on this was maximally likely short of explicit agreement, which was disallowed by the no-discussion rule.

Finally test 4, which was audio recorded, presented the group of five or six people with five slightly longer pieces with the instruction to agree a set of three appropriate words from the given list. The pressure here was to necessity of reaching agreement before the test proceeded to the next item. No time restrictions were imposed. The discussions lasted between five and fifteen minutes. This provided the author with tape-recorded evidence of the strategies adopted by a group of people negotiating their way to agreement.

4. Analysis of the Vocabulary

The objective of this analysis was to assign classes of use to descriptive words independently both from particular pieces of music and from other members of a test group. The experiment sought to analyse first the usage of the vocabulary to describe distinct categories of musical experience and second to assess the vocabulary in its capacity to convey a range of positive to negative evaluations. The researcher chose the former categories after discussion with professional musicians. These categories

² Sony DM6 Walkman Professional

were: value, for example the greatness of the piece; speed; mood, sad or happy etc.; tunefulness; and finally rhythm.

A 160-item questionnaire was constructed and issued to 58 volunteers. See Table 4 below. To make the analysis easier and the task of completion quicker the boxes only had to be ticked. Participants reported taking upwards of two hours to complete this, making the almost 100% return quite remarkable.

Word	A	B	C	D	E	F	G	H	I	J	K
category of word □	rhythm	tunefulness	mood	speed	value	don't know	very positive	positive	neutral	negative	very negative
sympathetic											
fluent											
forceful											
polished											
pastoral											
lacklustre											
light											

Table 4: The structure of the main vocabulary survey

Optical Mark Reading technology was used to create coded versions of responses. A small segment of the OMR output is reproduced below in Table 5.

sympathetic	fluent	forceful
nnnnnFnnnnn	AnnDnnnHnnn	AnCDnnnnHnnn
nnnnEnnnHnnn	nnnDnnnnHnnn	AnnnnnGnnnn
nBCnnnnHnnn	nBCnnnnnnInn	AnnDnnnHnnn
nnCnnnnHnnn	AnnDnnnHnnn	nnCnnnnHnnn
nnCDnnnHnnn	ABnnEnnnHnnn	AnCnnnnnnInn
nBCnEnnnHnnn	AnnDnnnnInn	nnnDnnnnnnK

Table 5: Sample OMR output

Any boxes ticked in Table 4 by each participant are reflected in Table 5 by an upper-case letter A to K with all other cells labelled with an “n” for Not filled. These spreadsheets were exported as comma delimited files into relational database management software so that SQL (*Sequel*) interrogations could be used allowing counting, alphabetic sorting, string chopping and fuzzy searching. Thus exact matches and, most crucially, similar patterns could be found using the SQL “like” function. Statistics can also be derived showing the extent and size of agreement. 308 different patterns were found amongst the total 9280 submitted. Many analyses became possible with this technique but the most interesting for this research was the ability to speedily find the words exhibiting the handful of most commonly occurring patterns.

5. The Second Group Experiment

Groups again met in a domestic environment and using paper records and cassette tape audio the activities were recorded for subsequent analysis. It has been suggested by Sloboda (1999) that the physical situation of experiments might affect the outcome. He notes that some experiments, carried

out in laboratories, may owe at least some part of their outcome to the artificiality of the surroundings. Given that the normal surrounding for listening to recorded music is the home, the author decided to apply the normalisation as far as possible by inviting the 19 participants to his home in three smaller groupings of 6, 6 and 7. This had the further benefit of allowing the use of high quality domestic, rather than lower quality institutional, playback equipment; the actual sound of the music was easier on the ears of the participants. In addition, since the study is specifically of informal discourse, the informality can be increased by the provision of tea, biscuits, wine, sandwiches etc. This is not a trivial issue. Most informal discussion of music takes place between concertgoers in bars during intervals or in bars after the concert. If anything the tea is the unreality.

For the experimenter the music to be played was listed along with a predicted word set as shown in table 6. This set was predicted after discussion with a colleague to act as a baseline for further analysis. As noted below it turned out to be no more than an experimental whim, like some sort of minority report.

Item Sequence	Music (circa 1 or 2 minutes)	Vocabulary Set
1	Prokofiev: Alexander Nevsky; The Battle on the Ice (opening) Decca CD 410 164-2 Track 1	icy, tense, crystalline, glacial, graphic
2	Milhaud: Scaramouche (3 rd movt.) DG LP 2531 389 Side 2 Track 3	vivacious, lively, joyous, rollicking, spirited
3	Weill: Surabaya Jonny : Happy End (opening) DG LP 2563 585, Side 1, Track 9	grieving, poignant, sentimental, passionate, theatrical

Table 6: Second Group Experiment: musical extracts and predicted associated vocabulary set (extract)

The participants were given only the word sets and no indication of the music chosen. All discussions were recorded on audiotape as previously noted. The data was collected subsequently from those tapes so as not to disturb the listening environment with pauses for the researcher to finish note taking.

It was predicted that with just 15 extracts and 15 lexical sets the likelihood of agreement was being maximised. Table 7 shows the way in which data was recorded for comparison. It also shows a little of the continued lack of agreement. The darker shading (in red on the original) highlights the final decisions of groups and the greyed out highlights show the word sets also discussed. This matrix consolidated the experimental record keeping in a way the author found useful for later communication with fellow researchers

Item Sequence	Music	Vocabulary Sets Discussed & Agreed														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Prokofiev: Alexander Nevsky; The Battle on the Ice (opening)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
The second group discussed 8 for sometime before deciding to agree on 11. All groups discussed 8 and 10 as possibilities but there was no final agreement. The predicted choice was 8.																
2	Milhaud: Scaramouche (3 rd movt.)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

All three groups discussed 3 but just two of the groups finally agreed on it as the prime descriptor set. The predicted choice was 7.

Item Sequence	Music	Vocabulary Sets Discussed & Agreed														
3	Weill: Surabaya Jonny (Happy End) (opening)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A rare agreement on the predicted choice, 9, and by all participants. Interestingly the second group never discussed any other set whereas the others ranged widely.																

Table 7: Second Group Experiment: agreed vocabulary set attachments (extract)

This final experiment sought to find agreement between people when the language was restricted and the music was kept within prescribed categories. The extent of disagreement was extreme with only two of the 15 extracts gaining agreement in all three groups and even then only one matched the predicted choice, making the “predicted choice” an irrelevance. The range of vocabulary discussed was very wide.

The author was led to the conclusion that, in essence, listeners do not agree in their predications, at least when a discrete vocabulary is imposed. It was discussion of this point that led to the theory being revised. Work is now ongoing that focuses on figurative, and therefore mainly phrasal structures. Initial results seem to suggest this is going to be more fruitful (Billinge and Addis 2003). The original aim of an artistic decision support system now seems more distant and possibly less interesting.

6. Methodological Conclusions

Earlier studies of the language of musical effect (for example: Gundlach 1935, Hevner 1936, Gabrielssohn 1973) were not clear about the procedures used to compile an initial descriptive vocabulary or about the approach taken to analysis of the corpus. The selection, mainly, of musically interested participants and the decision to have no control group might need consideration but the nature of the results did not imply that this approach was mistaken. The author believes that a control group would be unlikely to share the lexicon sufficiently to contribute. The focus of the research is on the means of communication. Those not sharing the language of “music talk” would fail to communicate and thus contribute no useful data. The decision to derive the initial lexical set from a mixture of personal knowledge and published sources was satisfactory because extensive user input allowed a means of refinement that gave an acceptable lexical set from the user’s viewpoint. The subsequent use of reduced and multi-element lexical sets provided subjects with a more common vocabulary and hinted at the need to extend this research into tropic communication where most figurative language is phrasal rather than lexically discrete. The use of a domestic environment for group meetings seemed to encourage verbal exchange in a way not reported by other researchers. The author has now accumulated many hours of natural conversation as well as substantial paper records. This corpus remains valuable despite a fairly drastic revision of theory (Billinge and Addis 2001, 2002a, 2002b). The tapes are currently being reanalysed to provide input data for experimental models of this inferential mode of discourse (Billinge and Addis 2003).

References

- Billinge, D. (2001). *An Analysis of the Communicability of Musical Predication* Unpublished PhD Thesis: University of Portsmouth.
- Billinge D. & Addis, T. (2001). *Some Fundamental Limits of Artistic Decision Making* Proceedings of the AISB’01 Symposium on Artificial Intelligence and Creativity in Arts and Science: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
- Billinge D. & Addis T. (2002a) *Modelling the Role of Metaphor in Artistic Description*: ECAI 15th European Conference on Artificial Intelligence, Lyon, France. Workshop 17 Creative Systems: Approaches to Creativity in AI and Cognitive Science pp 55-58
- Billinge D. & Addis T. (2002b) *Towards Constructing Emotional Landscapes with Music*: in George S. (Ed) *The Visual Perception of Music Notation* (US publication pending IGP 2003)

- Billinge D. & Addis T. (2003) *The Functioning of Tropic Communication: A Mechanism for Consistent Figurative Descriptions of Artistic Effect*. AISB'03 Symposium on AI and Creativity in Arts and Science
- Gabrielsson, A., (1973) *Adjective ratings and dimension analyses of auditory rhythm patterns*. Scandinavian Journal of Psychology 14, pp.244-260
- Gundlach, R.H. (1935) *Factors determining the characterization of musical phrases*. American Journal of Psychology 47, pp.624-43
- Hevner, K., (1936) *Experimental studies of the elements of expression in music*. American Journal of Psychology 48, 246-268
- Sloboda, John A (1999) *Everyday uses of music listening: a preliminary study*; in Yi, Suk Won (1999) *Music, Mind and Science*. Seoul National University Press. Korea
- Zipf, G (1949) *Human Behaviour and the Principle of Least Effort*. Addison Wesley, London