APPENDIX C

Preliminary Quantitative Results

In Chapter 10, I proposed that after additional improvements to Musicat, it would be feasible to compare it with Temperley's model from *The Cognition of Basic Musical Structures* (CBMS) (Temperley, 2001). Because the current version of Musicat is required to put group boundaries at measure boundaries (taking into account Musicat's shifting of entire melodies to account for pickup bars), it is impossible to compare the grouping structures generated by Musicat and the CBMS grouping program in a way that puts the two programs on equal footing. The Essen folksong corpus, which was used in Temperley's reported results from CBMS, includes many instances of group boundaries that are not found at measure boundaries (even if we account for pickup bars). Temperley's program does not have the measure-boundary restriction on grouping that Musicat does.

Because of this restriction in Musicat, I generated a simplified version of Temperlery's 65-song subset of the Essen corpus, in which all mid-measure group boundaries were "rounded" to occur at the start of the measure in which they occurred (after shifting all barlines sideways, if necessary, to account for pickup bars). For many melodies this has little effect; for others it makes Musicat's grouping boundaries more likely to line up with the ground truth. This simplification, to be sure, will artificially improve Musicat's reported

results to some degree, but I decided to include these preliminary results in any case, as a sneak preview of future testing to be done.

One other change was made to the ground truth file: several of the 65 melodies used by Temperley in testing CBMS were uninterpretable by Musicat, because they included notes of durations that Musicat could not handle. These melodies were excluded from this test; Musicat was tested on a subset of Temperley's subset of the Essen corpus.

I stress that these changes mean that the following comparison artificially inflates Musicat's results relative to those of CBMS, but perhaps not to a very large degree. Future testing, on even footing, is imperative, and will be carried out once I have improved Musicat and removed the group-boundary restriction.

The following table gives results obtained by running Musicat 20 times on each melody in the dataset, with a different random seed for each run. The boundaries indicated in the Essen corpus were converted into groups in a straightforward manner in order to generate a ground truth file, but no meta-groups were available, so all meta-groups generated by Musicat were ignored. I also included results for Musicat running on the Simple Melodies and Complex Melodies from chapters 6–7 of this thesis. Many of these melodies were used during the development of Musicat; this table shows results from testing on the training data for the Simple and Complex melody sets, and hence these values are also higher that they would be on a completely separate test set (Musicat was never run on the Essen subset during development, fortunately). For the Simple and Complex melodies, I have also included desired meta-groups and desired analogies in the ground truth.

I also report the number of groups and analogies that are "extra" for each run. The percentages in the table for extra groups or analogies were calculated by dividing the number of groups or analogies that were generated during a run but that did not exist in the ground truth by the total number of groups or analogies generated by the program for the same run.

	Groups correct	Groups extra	Analogies correct	Analogies extra
Musicat: Simple Melodies	83%	14%	48%	66%
Musicat: Complex Melodies	68%	43%	27%	78%
Musicat: Simplified Essen sub-subset	**74%**	39%	n/a	n/a
CBMS: Essen subset	**76%**	25%	n/a	n/a

Table 3: Preliminary results for Musicat, compared with reported results from CBMS.

Musicat does better on the simple melodies than on the complex melodies for all four metrics that were computed. Notice that many extra groups and analogies were generated, especially for the complex melodies. However, the scoring function I used treats groups and analogies in a binary fashion: a very weak group is weighted just as highly as a very strong group, so if very many extraneous weak structures are generated, they will have a large negative effect on the performance, even though those structures may not be very important to the program. A more fair test might use a threshold to remove weak structures from the calculation (just as is implemented in the user interface by the detail slider).

On the simplified subset of the Essen subset used by Temperley, Musicat generated 74% of the desired groups, trailing behind CBMS just slightly in performance. However, recall that these numbers are artificially inflated for Musicat. Even so, the numbers are encouraging to me, because it seems that Musicat is in striking distance of the accuracy of CMBS. Musicat and CMBS are both cognitively inspired models, but whereas CBMS is an off-line algorithm, Musicat generates groups in real-time, which is arguably more difficult. I was encouraged to see Musicat's 74% number, although a more fair comparison must be made in the future.