

Supplementary description of coding and modeling of events

To see how the modeling process works, we return to the example from John Coltrane’s ‘Equinox’ discussed in section 3.1. This song contains a VI7 chord on the downbeat of measure 9, but does *not* contain a III7 on the downbeat of measure 8, even though such a chord appears in other songs in the corpus (Miles Davis’s *Pfancin*, for instance). After the culling and analysis of the corpus described in the preceding sections, our data takes the form of a series of rows, each of which corresponds to some possible event in a particular position in a particular piece. Each column codes some piece of rhythmic, harmonic, or other type of information about the event in question. One final column codes whether the possible event in question did or did not occur in this particular position in this particular piece. So the rows for the two possible events in question in Equinox would look something like Figure 1.

(1) Database entries for the occurring VI7 chord and the non-occurring III7 chord in ‘Equinox’

Song	Comp	MetPos	Root	PreRMot	FolRMot	Attach	Embed	LDAttach	Occur
Equinox	JohCol	14	3	9	7	1	2	0	0
Equinox	JohCol	16	8	4	1	1	1	0	1

Note that some columns are omitted for ease of exposition. These entries encode various aspects of the possible harmonic events under consideration. The first two columns encode the song and composer associated with the events. The third column encodes the metrical position in which these possible events might have occurred (where 0 is the downbeat of measure 1 and 23 is the second half of measure 12). The ‘Root’ column shows the root of the chord in question, in semitones above tonic. ‘Pre’ and ‘Fol’ root motions encode the interval in descending semitones between the root of the chord in question and the root of the preceding and following occurring chords (respectively) in the piece. The ‘attach’ column encodes whether the event in question can be attached into the tree structure assigned to ‘Equinox’. The ‘Embed’ column encodes the number of nodes between the

chord in question, if it were attached into the tree structure, and the closest chord in the blues skeleton. ‘LDAttach’ encodes whether the possible attachment coded in the ‘Attach’ column is long-distance, that is, whether it involves a dependency between non-surface-adjacent chords. Finally, the ‘Occur’ column encodes whether the event in question occurred in the indicated metrical position in John Coltrane’s ‘Equinox’. The entire database would have a couple thousand rows like this, for all of the occurring and non-occurring possible chords in the corpus.

The job of the logistic models is to predict the probability of the final column being 1 (that is, the probability of chords occurring) using some of the information in the other columns. Different models use different collections of columns to make such predictions. All of the models under consideration here would have access to the fact that these events are being evaluated for the song ‘Equinox’. And they would generally find that ‘Equinox’ has fewer chords than the average piece in the corpus, so the probability of any chord occurring in Equinox is somewhat low, relative to other pieces.

The baseline models introduced in the next section attempt to predict the probability of chords occurring based entirely on non-syntactic information. So, for instance, the *root only* model would attempt to predict the probability of these two chords occurring given only the information in the ‘root’ column. It would assign the III7 chord around a 6% probability of occurring, and the VI7 chord around 9%. This is based on solely on the observation that III7 chords occur in about 6% of the positions where the corpus says they could occur, and VI7 chords in about 9% of the positions where the corpus says they could occur. The *position only* model ignores the harmonic properties of these chords entirely (even their roots), and attempts to predict their probability of occurring based only on the ‘MetPos’ column. This model estimates the probability of any given chord change occurring on the downbeat of measure 8 (position 14 here) around 11%, and the downbeat of measure 9 (position 16 here) around 17%.

More sophisticated models use syntactic information to try to predict chord occurrence. For instance, the *unigram* + *root-motion* model would use both overall root frequency (the ‘root’ column here) and the interval formed between the root of the preceding chord and the one in question (‘PreRMot’) to predict outcomes. This model would assign decrements of about 1.4 logits and 1.1 logits, respectively, to a III7 and VI7 chord occurring relative to a tonic chord. It would then assign 0.41 logits to the downward-5th motion formed by the III7 chord and the following VI7 chord, and -0.19 logits to the descending semitone motion formed by the VI7 chord and the following V chord. The final probabilities this model assigns to the two events are about 17% for the III7 chord and 30% for the VI7.

Finally, a CFG-based model would use the tree-based columns to try to predict chord occurrence. For instance, a depth-of-embedding + root model would use the information from the ‘root’ column and the ‘embed’ column to try to predict the ‘occur’ column. This model would assign decrements of 0.92 logits to a III chord and 0.55 logits to a VI chord, relative to the tonic. It would assign decrements of 0.87 logits to a chord embedded one level below the skeleton and 1.46 to a chord embedded 2 levels below, relative to a chord contained in the skeleton. The final probabilities this model would assign to the two events are about 9% for the III7 chord and 21% for the VI7 chord.

With these estimates in place, we can compare the likelihood each model assigns to the miniature corpus consisting of the two events in question. It is simply the joint probability of the VI7 occurring and the III7 not occurring. This is shown in figure 2.

(2) Likelihood of the two events in the ‘Equinox’ example

Model	p (i)	p (j)	p (i & ~j)
Root	0.09	0.06	0.085
MetPos	0.17	0.11	0.151
Uni+RM	0.30	0.17	0.249
Uni+DOE	0.21	0.09	0.191

It can be seen in figure 2 that the more complex models do better on this mini-corpus than the baseline models: they assign higher probabilities to the observed outcomes (VI occurring and III not occurring). These models, however, also use more parameters. To simplify a bit, the baseline models use only 1 kind of information to predict outcomes while the complex models use 2 kinds of information (the real models in fact use one parameter for each separate root, root-motion, metrical position, depth-of-embedding, etc.). The BIC is a way of trading off this complexity against the gains in terms of likelihood (the AIC is a slightly different way of computing the tradeoff but uses the exact same information). And of course, the real models are assessed not against these two events from ‘Equinox’ but against the entire database of occurring and non-occurring possible harmonic events. But the overarching logic is aptly illustrated by this example.