Weighing conflicting constraints: A maxent approach to textsetting

Bruce Hayes
Department of Linguistics
UCLA

Some core ideas of Lerdahl and Jackendoff (1983)

- Their theory generates **structural analyses**, intended as psychologically-real representations for how music is apprehended by people.
- These analyses are obtained by selecting from a set of logical possibilities, determined by the **well-formedness rules**.
- The selection is made according to a set of **preference rules**.
- Preference rules can **conflict**, resulting in vague or ambiguous perceptions.

A crucial issue left unaddressed in L+J

- The theory is **underformalized** — it cannot
  - make numerical predictions
  - be rigorously tested with corpus or experimental data
- Hence LJ emphasize (persuasive) particular examples.
- **Comment**: LJ were brave to do this, and it was worth it.
  - Conceptualization is at least as important as formal implementation.
  - They gave us a nice research problem — how to formalize the theory?

So why didn’t LJ formalize?

- They explain this very clearly (see “Remarks on Formalism,” pp. 54-55). Two reasons:

I. The Gradience Problem

- People's judgments about the perceived structure are often ambiguous, or not clear-cut.

> - "[Our] rules fail to produce a definitive analysis [because] we have not completely characterized what happens when two preference rules come into conflict."
> - [Numerical schemes, like rule weighting] "allow only positive and negative judgments; not ambiguous or vague ones."

II. The "Apples and Oranges" Problem

- How to assign weights to preference rules of utterly different types? E.g.:

> "How much local instability in grouping, or loss of parallelism, is one to tolerate in order to produce more favorable results in the reductions?" (p. 54)

Scrolling through 25 years of history

- Music cognition has flourished, by using
  - theory
  - data corpora
  - experimentation
  - computational modeling

David Temperley’s modeling program

I. The Cognition of Basic Musical Structures (2001)

- Formalizes preference rules (using weights, as L+J suggest), and succeeds in explicitly modeling lots of data. **But**:
  - No principled basis for assigning the weights; they were “mostly set by trial and error”.
  - Can’t predict gradience.

II. Music and Probability (2007)

- Temperley abandons preference rules, adopting instead an eclectic mix of **probabilistic** models.
  - ➢ Again he addresses various data domains, and gets good modeling results—this time including gradience.

Could there be a probabilistic implementation of preference rules?

- My goal is to show that this is possible.
- It also seems desirable:
    - Preference rules embody the theory at a highly abstract level, as in the “computational theory” of Marr (1983).
    - Their content is fully accessible to human understanding, which should aid progress.

Two premises

- **Premise 1**: preference rules are weighted, and the weights are **learned** by people when exposed to idiom-specific data.
  - I conjecture that this is the solution to the apples/oranges problem—you learn to balance apples and oranges as they are balanced in the musical idiom you are learning.

- **Premise 2**: Certain mathematical tools, newly developed by computer scientists, provide a suitable formalization for gradiently-operating preference rules.

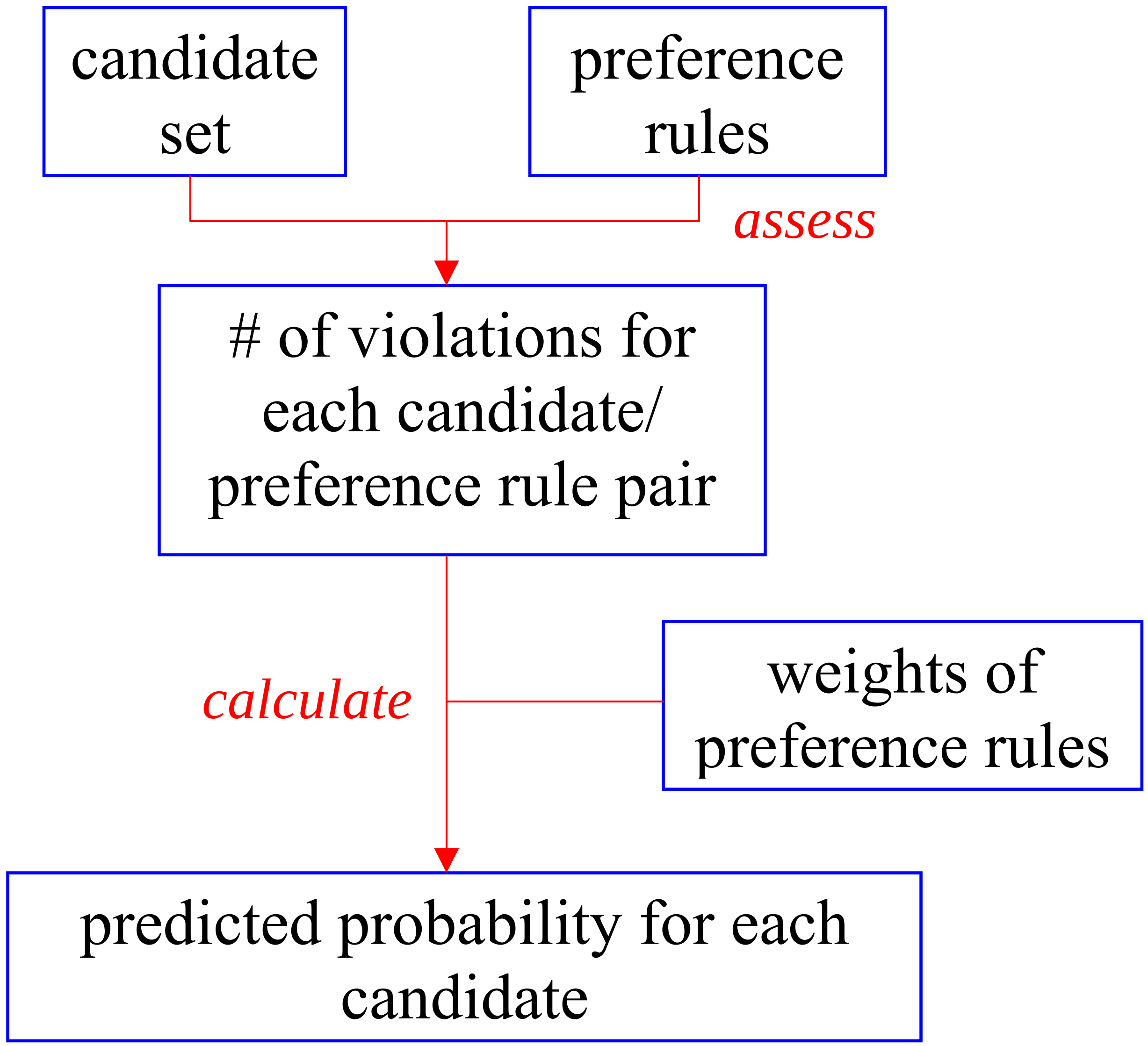
Rest of the talk

- Describe **maximum entropy (maxent) grammars** and their associated learning algorithm.
- Describe why they are a good candidate for a formal implementation of gradient preference rule theory.
- Case study: the "textsetting problem" (Halle and Lerdahl 1993).

Maximum entropy grammars: starting point

- In some domain of analysis, assume a **candidate set**.
  - E.g. every possible Grouping Structure (L+J) for a passage of music.
- Each preference rule is assigned a numerical **weight**.
- Each preference rule assigns **violations** to candidates, denoting imperfection, following some formal scheme created by the analyst.

Maxent grammar: outline model



The probability calculation

1. For each candidate, find the **dot product** of weights and violations (sum of individual products) over the set of preference rules.
2. Take *e* (≈ 2.718) to the result.
3. Do the same for all candidates and sum overall, forming a value termed Z.
4. Probability of a candidate = its share of Z.

Finding the right weights

- Assuming a training set (e.g., a large body of music in a particular idiom)
- Weights are set to achieve an objective: *maximize the predicted probability of the data in the training set, given the set of preference rules.*
- … thus minimizing the predicted probability of what is *not* in the training set.

- The predicted probability of the data is calculable (as a simple product).
- So finding the best weights becomes a mathematically well-defined search problem.

Searching for the best set of weights

- No time to cover here, but I note that the relevant algorithms are
  - proven to converge
  - fast enough for the project to be feasible
- For extensive discussion and references, please consult
  - Hayes, Bruce and Colin Wilson (in press) "A maximum entropy model of phonotactics and phonotactic learning," to appear in *Linguistic Inquiry*.

Case study: the textsetting problem

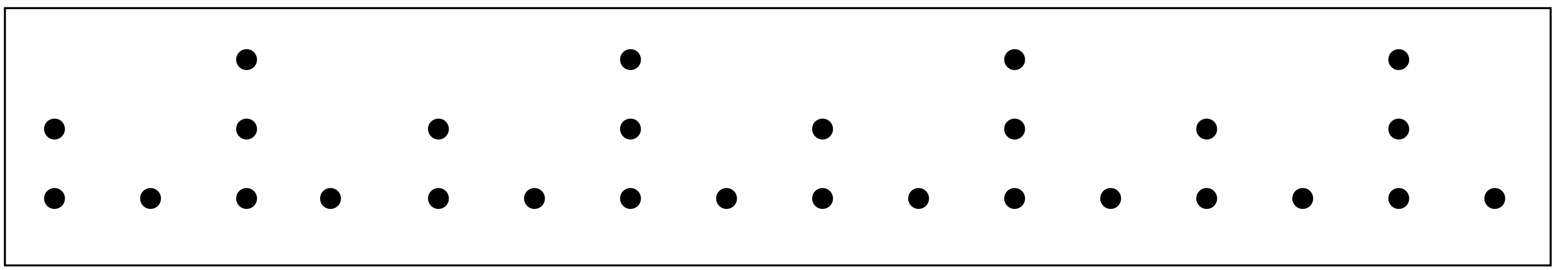
- When we learn the words of a novel verse of a song, how do we line then up against the song's rhythm?
- People know how to do this, and agree fairly well in their intuitions of preferred alignments.

Example

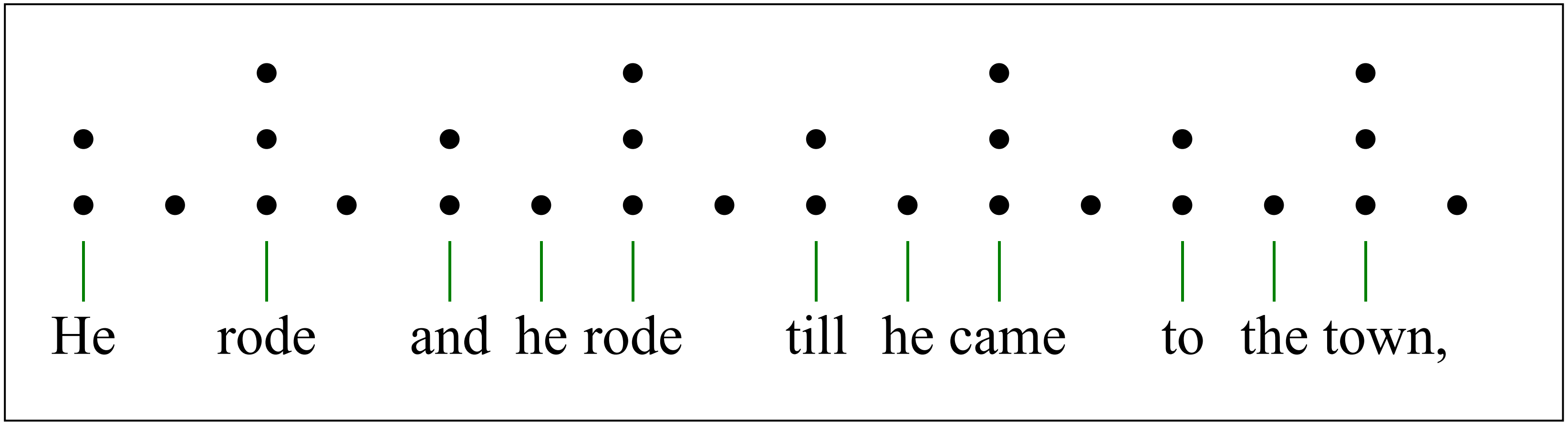
Assume this text:

He rode and he rode till he came to the town,

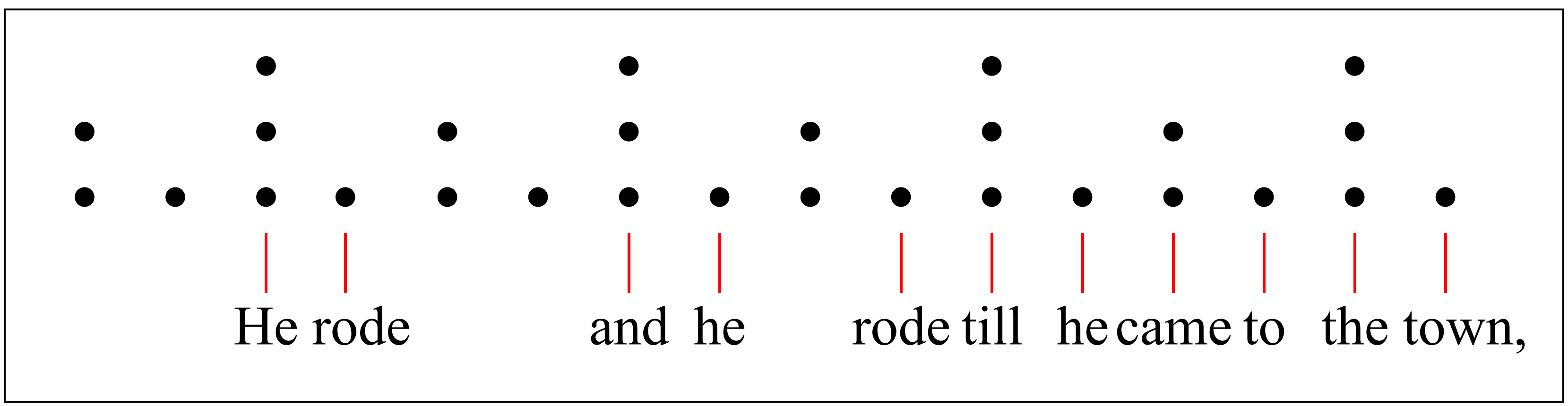
and a L+J-style grid for a single line of this song:



We must predict:



and not bad alternatives like:



Gradience

- People often find multiple settings to be ok, varying along a continuum of acceptability.

Earlier work on the textsetting problem

- Dell (1975, 2004)
- Stein and Gill (1980)
- Oehrle (1989)
- Halle and Lerdahl (1993); Halle (1999, 2004)
- Hayes and Kaun (1996)
- Hayes (in press)
- Keshet (2006 ms.)

Preference rules applied to textsetting: a minor difference

- Production, not perception:
  - Which of the (several thousand) alignments of syllables to grid does the speaker prefer?

Data to be modeled

- Hayes and Kaun (1996): 9 consultants each chanted the text of 670 lines of traditional English folk song, in rhythm.
- Goal is to model the **share of the vote** that each setting got—this will serve as an approximation for gradient intuition.

Preference rules employed

- You’re going to have to take these mostly on faith …
- Many are identifiable as restatements, or contextually applicable versions, of preference rules in L+J.
- Others are related to how language is used to manifest rhythm—
  - This is the field of **metrics**, which has mostly worked with data from written verse.

Sample research findings in metrics

Stressed + stressless demands to match the grid more strongly if the two syllables are in the **same word**.

Stressless + stressed demands to match the grid more strongly if the two syllables are **at the end of a major phonological phrase**.

- Preference rules are included here to capture these effects.
- References: Halle and Keyser (1966, 1971), Kiparsky (1975, 1977), Hayes (1983, 1989)

Preference rules used

FILL S(TRONG BEAT)	implement L+J's MPR 3 (EVENT)
DON'T FILL W(EAK BEAT)	
FILL M(EDIUM BEAT)	
MATCH PHRASE-FINAL	implement MPR 4 (STRESS)
LEXICAL STRESS	
RISING LEXICAL STRESS	
*STRESS IN M	
*STRESS IN W	
REGULATE SW	implements both MPR 3 and 4
REGULATE MW	
REGULATE SM	
STRONG IS LONG	close to MPR 5 (LENGTH)

<table>
<tr><td>DON’T FILL 16</td><td rowspan="2">implements GPR 2 (PROXIMITY)</td></tr>
<tr><td>DON’T FILL 1</td></tr>
<tr><td>RESOLUTION</td><td rowspan="3">text-grid duration matching</td></tr>
<tr><td>AVOID LAPSE</td></tr>
<tr><td>WEAK RESOLUTION</td></tr>
</table>

An implementational issue

- To keep computation size reasonable, I took two very powerful preference rules:
  - FILL STRONG (“the strongest metrical positions must be filled with a syllable)
  - REGULATE SW (“don’t put stronger stress in W than in an adjacent S”)

and gave them the status of Well-formedness rules, thus limiting the candidate set.

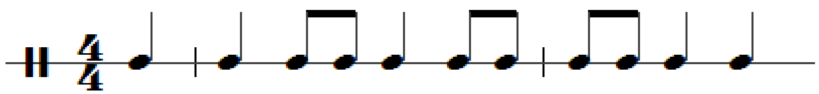
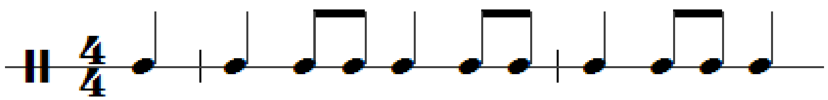
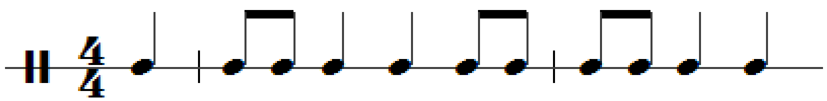
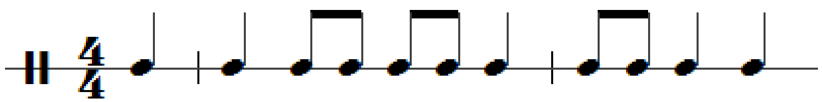
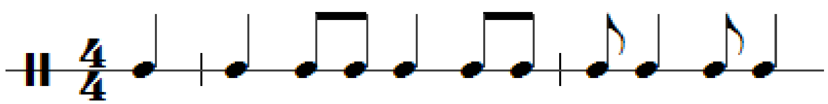
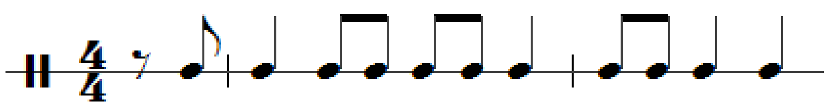
The simulation summarized

425	lines (removed lines found only in some stanza types)
8.4	average # valid "votes" per line / 9
2.2	average # of distinct settings among the votes
117	Average # of candidates

- **Goal**: find weights that predict the distribution of votes as accurately as possible
- I also did "cross-training" runs: train on one half, test on other; this yielded similar results.
- I used maxent software created by Colin Wilson.

Results I: sample output

‘Come all that’s around me and listen awhile’

Setting	Votes	Pred. score
	5	0.460
	1	0.155
	0	0.117
	1	0.117
(others, getting no votes)		…
	1	0.0038
	1	0.0025

Results II: Raw correlation

- For the entire set of candidates, the correlation r of predicted probability vs. "vote share" is r = **0.883**.
- This is only a rough measure, since most values for both voting and prediction are at or close to zero.

Results III: Data and predictions in bins

Predicted probability

Vote share

	0-.1	**.1-.2**	**.2-.3**	**.3-.4**	**.4-.5**	**.5-.6**	**.6-.7**	**.7-.8**	**.8-.9**	**.9-1**
0-.1	48462	191	41	10	7	3	1			
.1-.2	259	34	19	4	3	3	2	1	1	
.2-.3	67	13	10	4	2	2	5		1	1
.3-.4	26	12	11	1	4	2	4	3	3	
.4-.5	12	13	6	3	6	3	2	4	4	
.5-.6	6	6	8	4	8	3	7	3	7	
.6-.7	3	1	5	5	3	6	17	6	14	1
.7-.8	4	5	2	4	4	6	12	6	18	1
.8-.9	2	4		4	3	12	20	13	33	5
.9-1		2	1	2	4	9	28	24	27	12

Improvements possible?

- Preference rules could be improved, I think.
- Keshet (2006), working non-gradiently, has discovered some new and interesting rules, but I've not had time yet to implement them.

Differences between consultants

- Hypothesis: the set of preference rules embodies the general theory, part of the competence of all participants (cf. L+J, 96).
- Individual idiosyncrasies must be due to consultant-specific weighting.
- We can detect this by *training the weights on the data specific to each consultant*.

Example: RH vs. DS's weights for two preference rules

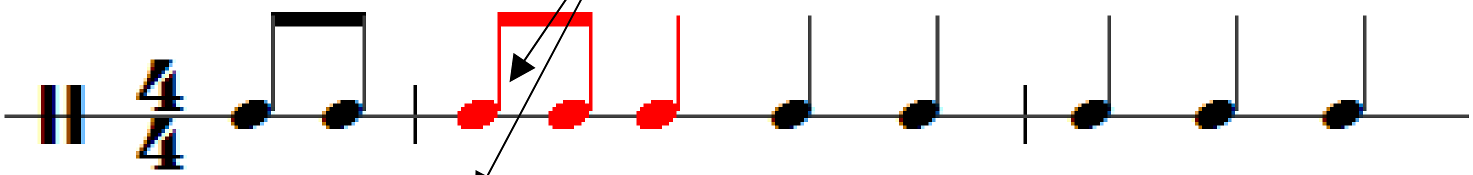
	RESOLUTION	STRONG IS LONG
RH	1.472	3.418
DS	2.480	0.879

- RESOLUTION (Kiparsky 1977, Hansen 1990, Hayes and Kaun 1996: Render as short any stressed syllable that is not word-final.
- STRONG IS LONG (≈ L+J, MPR 5)

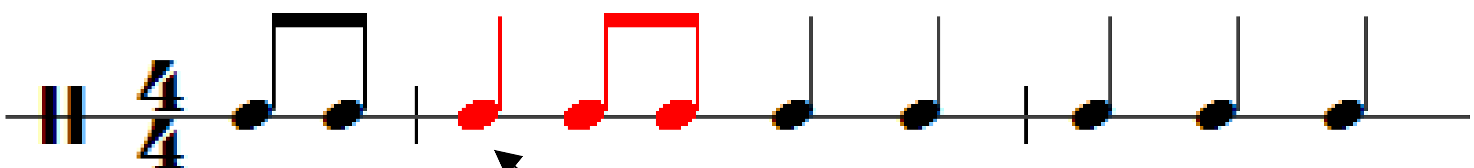
These different weights predict different behavior.

“The remarkable day that I was wed”

Consultant DS’s setting satisfies RESOLUTION:



The re- mar ka ble day that I was wed



The re- mar ka ble day that I was wed

Consultant RH’s setting satisfies STRONG IS LONG.

DS and RH's own grammars predict these settings as favorites

Probabilities:

	RH's choice	DS's choice
RH's grammar	0.689	0.065
DS's grammar	0.251	0.819

Upshot

- The maxent approach not only characterizes the data as a whole fairly well, but gives us a means of characterizing individual differences in style.

Caveat: do RH and DS really have different grammars?

- Maybe, but my guess is that they are construing the experimental situation differently:
  - Each commands a variety of idioms.
  - They accessed different ones in performing the experimental task.

Summary

- The maxent approach shows promise, I think:
  - Solving the gradience and apples/oranges problems
  - Retaining the generality and interpretability of the preference rule approach.
- It's easy to apply, and if you would like to try it, I will gladly share the software with you (email next page).

Thank you

Author's contact information:

Bruce Hayes
http://www.linguistics.ucla.edu/people/hayes/
bhayes@humnet.ucla.edu

References

Dell, François (1975). Concordances rythmiques entre la musique et les paroles dans le chant: l'accent de l'e muet dans la chanson française. In Le soucie des apparences, Marc Dominicy (ed.), 121-136. Brussels: Editions de l'Université de Bruxelles.

Dell, François (2004) Singing as counting syllables: text-to-tune alignment in traditional French songs. Ms.

Halle, John and Fred Lerdahl (1993) “A generative textsetting model,” *Current Musicology* 55:3-23.

Halle, John (1999) *A Grammar of Improvised Textsetting*. Ph.D. dissertation, Columbia University.

Halle, John (2003) Constituency matching in metrical texts. Submitted for publication in the proceedings of the conference Words and Music, University of Missouri-Columbia, March 14, 2003.

Halle, Morris and S. Jay Keyser (1966) “Chaucer and the theory of prosody,” *College English* 28: 187-219.

Halle, Morris and S. Jay Keyser (1971) *English stress: Its form, its growth, and its role in verse*. New York: Harper and Row.
Hanson, Kristin (1990) Resolution in Modern Meters, Ph.D. dissertation, University of California, Berkeley.
Hayes, Bruce (1983) “A grid-based theory of English meter,” *Linguistic Inquiry* 14, 357-393.
Hayes, Bruce (1989) “The prosodic hierarchy in meter,” in Paul Kiparsky and Gilbert Youmans, eds., *Rhythm and Meter*, Academic Press, Orlando, FL, pp. 201-260.
Hayes, Bruce (in press) “Textsetting as constraint conflict,” to appear in Aroui, Jean-Louis and Andy Arleo, eds. (forthcoming) *Towards a Typology of Poetic Forms*. Amsterdam, Elsevier.
Hayes, Bruce and Margaret MacEachern (1998) “Quatrain form in English folk verse,” *Language* 74, 473-507.
Hayes, Bruce and Abigail Kaun (1996) “The role of phonological phrasing in sung and chanted verse,” *The Linguistic Review* 13, 243-303.
Keshet, Erza (ms., 2006) “Relatively Optimal Textsetting,” http://web.mit.edu/ekeshet/www/
Kiparsky, Paul (1975) “Stress, Syntax, and Meter,” *Language* 51:576-616

Kiparsky, Paul (1977) “The rhythmic structure of English verse,” *Linguistic Inquiry* 8.189-247.
Lerdahl, Fred and Ray Jackendoff (1983) *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.
Marr David (1983) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman.
Oehrle, Richard (1989). Temporal structures in verse design. In Paul Kiparsky and Gilbert Youmans, eds., *Rhythm and Meter*, Academic Press, Orlando, FL, pp. 87-119.
Stein, David, and David Gil (1980) Prosodic structures and prosodic markers. *Theoretical Linguistics* 7. 173-240
Temperley, David (2001) *The Cognition of Basic Musical Structures*. MIT Press, Cambridge.
Temperley, David (2006) *Music and Probability*, MIT Press, Cambridge.