

# Maxent grammars for the metrics of Shakespeare and Milton

Bruce Hayes (UCLA)

Anne Shisko (UCLA)

Colin Wilson (Johns Hopkins University)

LSA Baltimore

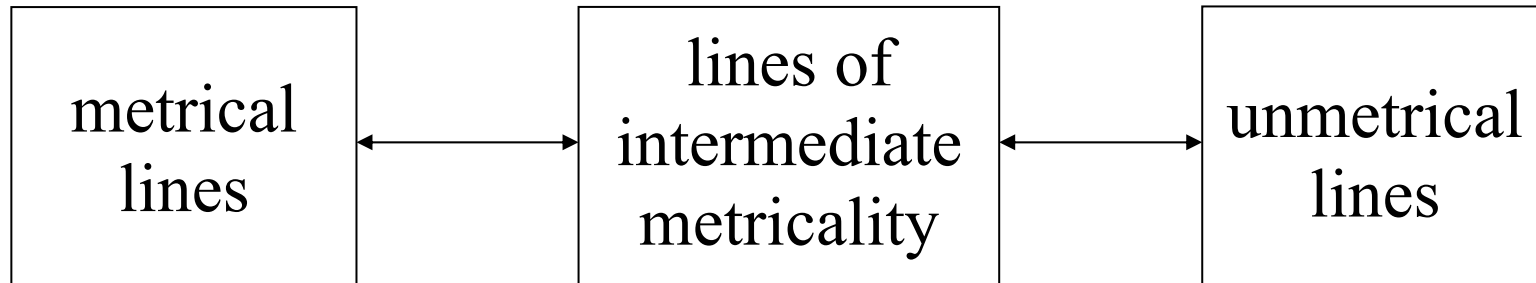
1/8/2010

# Background

- We adopt the theory of **generative metrics** (Halle and Keyser (1969, 1971) et seq.)
- Goal: to characterize with formal grammars the system guiding the poet's craft in creating lines of verse.

# Halle and Keyser's assumptions

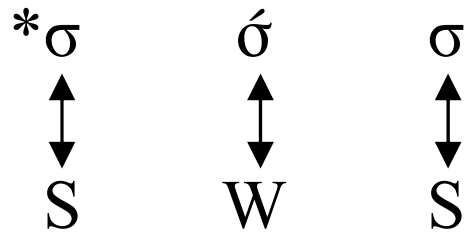
- There is in effect a continuum of well-formedness.



- We can describe this continuum with a quantitative model.
- Example: their quantitative model of line types in *Beowulf* (HK 1971, 147-164), with good qualitative fit to data

# Pitfalls in developing quantitative models in metrics — an example

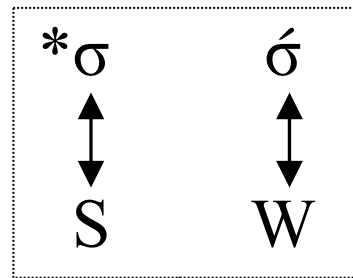
- A serious metrical grammar often has many constraints that overlap in their effects.
- Example: the classical **Stress Maximum Constraint** (Halle and Keyser 1969 et seq.), schematically:



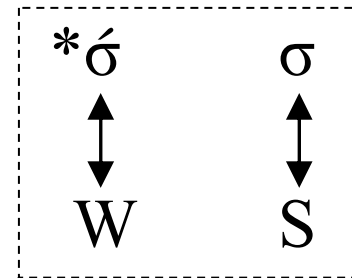
- This overlaps with constraints proposed earlier by Otto Jespersen (1901).

# Jespersen I + Jespersen II = Stress Maximum

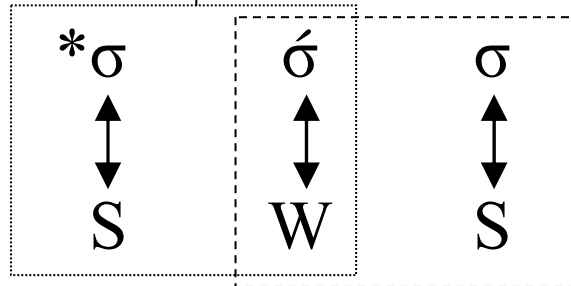
Jespersen I:



Jespersen II:



Stress Maximum Constraint:



# The fallacy of expected values

- If an outcome is absent or extremely rare, it might not mean anything — what if it is the **intersection** of two outcomes that we independently expect to be rare?

# Example

- Here is a rare outcome that is arguably meaningless.
  - Suppose: “*no speaker of Menominee stutters*”
  - Does the structure of Menominee miraculously prevent its speakers from stuttering?
  - Not likely: the U.S. stuttering rate is about 1%, and there are only 40 remaining Menominee speakers (< *Ethnologue*).
  - **Expected value** for the number of Menominee speakers who stutter = 0.4, so we learn nothing from the zero frequency.
- Metrical studies must make sure they do not fall into this trap.

# Avoiding the fallacy of expected values

- How can we know which constraints are doing the real work of the grammar?
- We need a system that rationally assesses constraints that overlap, in a quantitative way.
- Recent progress in mathematical linguistics suggests a way.

# Maxent grammars

- Recent work that has made use of maxent grammars in linguistics:
  - Goldwater and Johnson (2003)
  - Wilson (2006)
  - Hayes and Wilson (2008)
  - Hayes, Zuraw, Siptár and Londe (in press, *Lg.*)

# Maxent grammars — a rough sketch

- They are **constraint based** — like most work in metrics.
- Every constraint has a **weight**, a nonnegative real number
- Every form (e.g., line of iambic pentameter) is assigned a **probability**, using a standard formula.

$$\frac{\exp(-\sum_{i=1}^N w_i C_i(L))}{\sum_{x \in \Omega} \exp(-\sum_{i=1}^N w_i C_i(x))}$$

- $w_i$  = weight of  $i$ th constraint
- $C_i(L)$  = violation counts for line  $L$ ,  $i$ th constraint
- $N$  = number of constraints
- $\Omega$  = set of all logically possible lines

# Maxent grammars — finding the right weights

- Weights are established by **fitting against the data** (i.e. a training corpus) with an machine-implemented algorithm.
  - Data need not be input-output pairs, but can be simply a list of observed forms (Hayes and Wilson 2008)—here, a corpus of scanned lines of verse.

# Maxent grammars — goal in setting the weights

- Follow the principle of **maximum likelihood estimation**:
  - Set the weights so that the **combined predicted probability of the observed data** is maximized.
  - Since probabilities sum to 1, this means that the combined predicted probability of the unobserved data is minimized, making the grammar as restrictive as possible.

# Maxent grammars — two virtues

- Unlike similar current models, the math behind the weight-setting is **backed by proof** (Berger et al. 1996, Della Pietra et al. 1997).
- The system also permits **statistical significance testing** of constraints — see below.

# Empirical work — strategy

- Collect and annotate **corpora of verse**
- Compile a **constraint inventory** from the research literature in metrics
- Form maxent grammars by **weighting** these constraints
- This will help reveal which of the constraints are really doing the work in characterizing the metrical lines.

# Data corpus

- We selected two corpora of iambic pentameter
  - Shakespeare, *The Sonnets*
  - Milton, *Paradise Lost*, books 9-10.
  - both about 2000 lines
- Reason: both have been meticulously studied by metrists over many decades.

# Phonological annotation of the corpus I: stress

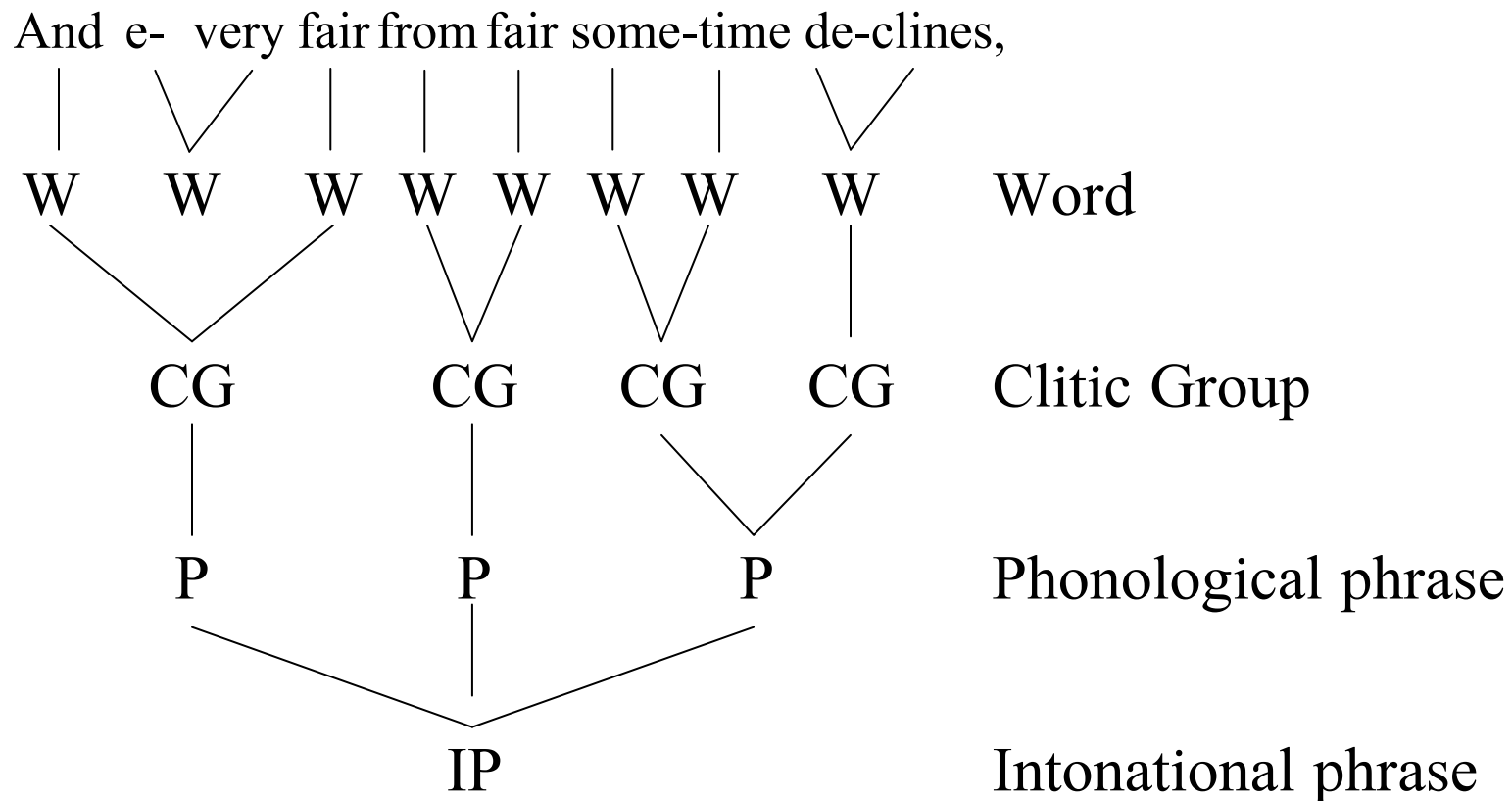
- We coded every line for stress: integers 1-4, roughly following Chomsky and Halle (1968)

1 2 1 4 1 3 3 2 1 4  
*And e- very fair from fair some- time de- clines,*

(Shakespeare, Sonnet 18)

# Corpus annotation II: phrasing

- We used the **Prosodic Hierarchy** (Selkirk 1982, 1986), using the categories and rules from Hayes (1989)



# Independent coders (Hayes/Shisko) get reasonably good agreement

	<i>Stress</i>			<i>Phrasing</i>	
same	20479	0.890	same	21457	0.933
1 off	2092	0.091	1 off	1406	0.061
2 off	377	0.016	2 off	120	0.005
3 off	55	0.002	3 off	11	0.000
<i>total</i>	<i>23003</i>		4 off	9	0.000
			<i>total</i>	<i>23003</i>	

# Constraints

- We did our best to cover the entire literature on metrics.
- We started with 39 constraints, taken from (among others), Jespersen (1901), Halle and Keyser (1969, 1971), Magnuson and Ryder (1970), Kiparsky (1975, 1977), Hayes (1983, 1989), Youmans (1989), Kiparsky and Hansen (1996), and Fabb and Halle (2008).
- We rendered them machine-readable using an ad hoc feature system.

# Software

- It does this:
  - reads a data corpus and constraint inventory
  - assigns each constraint a weight using an algorithm that implements the maxent principles
  - performs a statistical test on each weight.

# Results for Shakespeare (Stage I)

- Of 39 constraints, 17 received **zero** as weights.
- Meaning in maxent terms: *they play no role in the optimum explanation of the data.*
- 22 constraints remain.

## Results, Stage II: significance testing

- It is not sufficient to show that a constraint has a positive weight.
- We need to show that the observed positive weight is unlikely to have arisen through chance fluctuations in the data.
- The test we used: a **likelihood ratio test**:

$$2 * \log\left(\frac{\text{data prob. with constraint}}{\text{data prob. without constraint}}\right)$$

- This is distributed as a chi-square with one degree of freedom and yields the probability of the hypothesis “improvement could be true by accident”.

# Results of significance testing (Shakespeare): more constraints drop out

- Of the 22 remaining constraints, 4 failed to reach significance by the likelihood ratio test ( $p \approx 1$  for all of them).

# What emerged from our analysis?

- The constraints that do the work tend to be the ones that are **independent**, doing their work mostly on their own — but not always.
- Such constraints come in families:
  - matching the stress to the meter
  - matching bracketing: linguistic phrasal boundaries should line up with foot and line boundaries
- We will cover the rarer case of valid conjoined constraints later on.

# Match stress to the meter

- Overview: stress contour is matched to the meter, particularly so within simplex words and in the last foot.
- Details on next slide.

<i>Constraint</i>	<i>weight</i>	<i>Source</i>
Don't fill W with stress.	1.914	Hanson and Kiparsky (1996)
*Rising stress in SW	0.791	Magnuson and Ryder (1970)
*Rising stress in SW within a simplex word	3.357	Kiparsky (1975)
4 constraints: *falling stress in WS, unless major phonological break precedes (variation: level of break specified; limited or not to simplex words)	0.881 0.032 1.559 0.936	Kiparsky (1975). Grammar predicts less mismatch after weaker phrase breaks—Hayes (1989).
Rising stress is required in the final foot	1.442	Youmans (1989)

# Match phrasing I: don't interrupt lines, or feet, with salient phonological phrase breaks

<i>Constraint</i>	<i>Weight</i>
*P-phrase break inside a foot	0.720
*I-phrase break inside a line	0.027
*I-phrase break inside the first, or last, foot	4.521, 2.779
*Clitic-group break inside last foot	0.291

# Match phrasing II: demarcate the line boundaries with salient phrasal breaks

i.e. avoid enjambment

<i>Constraint</i>	<i>Weight</i>
Line boundary must coincide <i>at least</i> with a Clitic Group boundary	2.560
Line boundary must coincide <i>at least</i> with an Intonational Phrase boundary	3.119

# Constraints whose effects are deducible from simpler constraints

- We first consider Kiparsky's (1977) constraints that ban simultaneous **bracketing** and **stress mismatches**.
- Kiparsky's example: unmetricality of Wyatt's

For good is **the** **lífe**, ]<sub>IP</sub> ending faithfully

[W S]<sub>foot</sub> [W **S**]<sub>foot</sub> [**W** S]<sub>foot</sub> [W S]<sub>foot</sub> [W S]<sub>foot</sub>

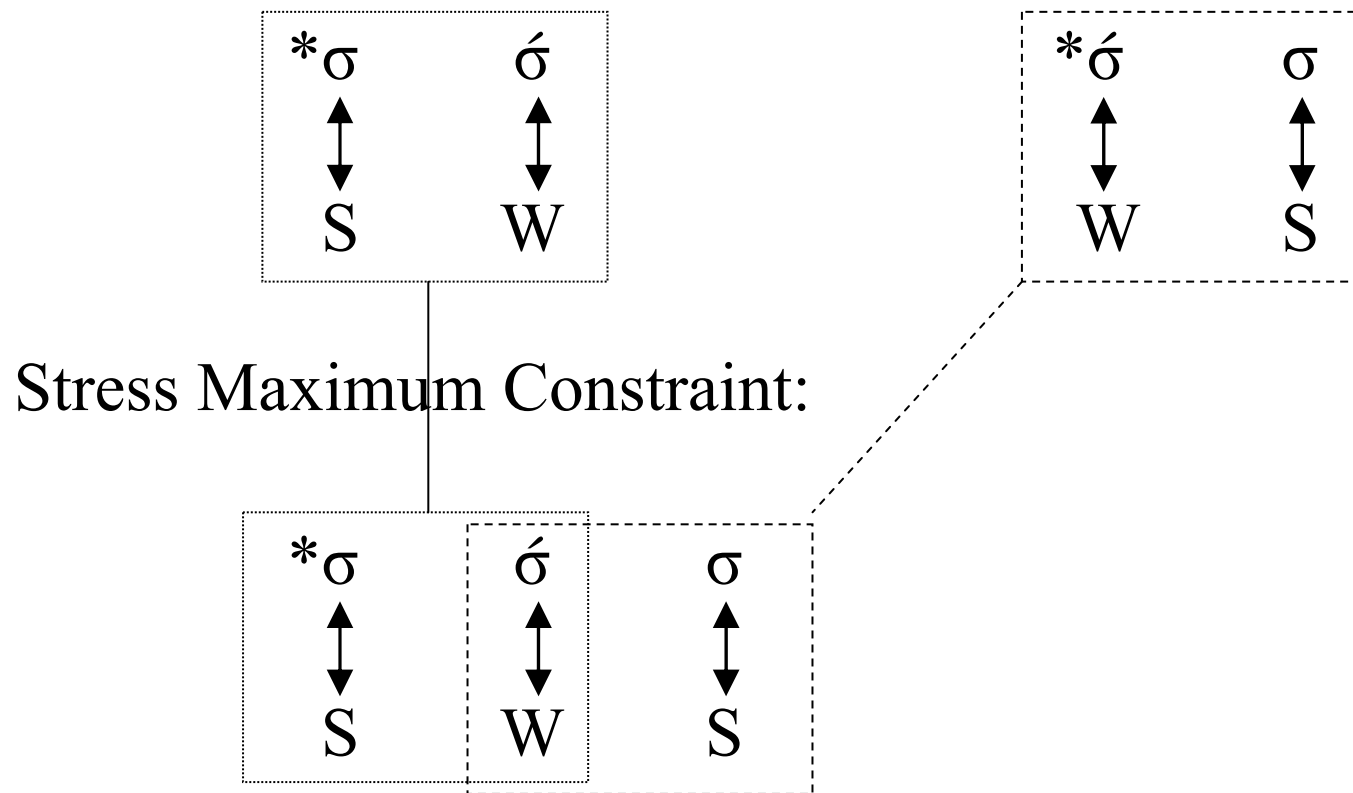
- We know (see above) that both parts of this constraint (no break within foot, no stress mismatch) are independently needed — fallacy of expected values?

# Kiparsky's constraints emerge as significant

<i>Constraint</i>	<i>Weight</i>
a. Don't place a rising stress sequence in SW at the end of an Intonational Phrase.	2.435
b. Don't place <i>stressless</i> + <i>stressed</i> in SW at the end of an Intonational Phrase.	0.880

# Stress Maximum Principle constraints also are deducible from simpler constraints

- Repeating the figure from before:



Are these also independently necessary?

# Results on various Stress Maximum constraints

- The following receive zero weight at Stage I of simulations:

Stress maximum in W (general)	Halle and Keyser (1969, 1971)
Stress maximum in W, all three syllables within the same IP.	
Stress maximum in W, all three syllables within the same PP.	

- The following constraint receives a positive weight (0.155), but does not pass the statistical significance test ( $p = 1$ ):

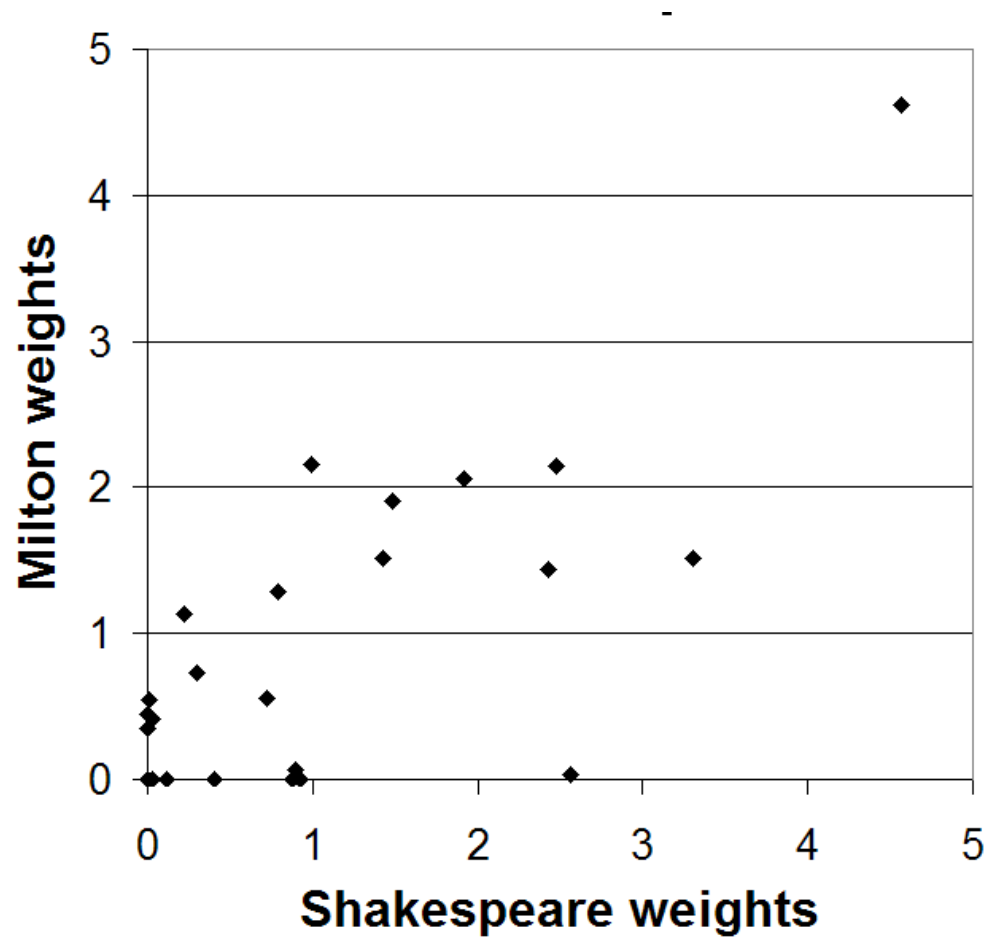
Stress maximum in W, where one of the neighboring syllables is in the same simplex word.	Fabb and Halle (2008)
--	-----------------------

# Caveats

- Our statistical test does not “confirm the nonvalidity” of a constraint; it only tells us that it is not supported by the data under consideration.
- These are quite recent calculations and need rechecking.

# What about Milton?

- Results are basically similar—scattergram of 39 weights:



# The “dialect” differences between Shakespeare and Milton claimed by Kiparsky (1975, 1977) are largely confirmed

- Details and weights omitted for lack of time; but they involve:
  - Within-word falling stress not after break
  - Whether stress on the first syllable of a mismatched rising phrase-final sequence matters
  - Demarcation of line breaks with phrase breaks—Milton far less strict

# Gradience-by-phrase-rank (Hayes 1989) is partially confirmed

- Hayes (1989) suggests a major source of gradience in metrics lies in the hierarchy of phonological phrasing. Constraints relying on phrase breaks get stricter the higher the rank of the phrase; licenses get more liberal.
- This shows up —to a limited degree—in the constraints and weights.

# Conclusions

- The maxent approach provides a formally rigorous way of stating and testing gradient metrical grammars, in which the constraints vary in their strength.
- By using this method, we can safeguard our work from the fallacy of expected values.
- Tentative simulation results: constraints proposed by metrists working on English iambic pentameter are confirmed in outline by this work, with the important exception of the constraints of the Stress Maximum family.

# Thank you

*Author's addresses etc.:*

Hayes/Shisko

Department of Linguistics, UCLA

<http://www.linguistics.ucla.edu/people/hayes/>  
[bhayes@humnet.ucla.edu](mailto:bhayes@humnet.ucla.edu)

Wilson

Department of Cognitive Science

Johns Hopkins University

<http://web.jhu.edu/cogsci/people/faculty/Wilson/>  
[colin@cogsci.jhu.edu](mailto:colin@cogsci.jhu.edu)

# References

- Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22:39–71.
- Chomsky, Noam and Morris Halle (1968) *The Sound Pattern of English*. New York: Harper and Row.
- Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:380–393.
- Fabb, Nigel and Morris Halle (2008) *Meter in Poetry: A New Theory*. Cambridge: Cambridge University Press.
- Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.
- Halle, Morris and S. Jay Keyser (1969) Chaucer and the study 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*, Harper and Row, New York.

- Hanson, Kristin and Paul Kiparsky (1996) A parametric theory of poetic meter. *Language* 72: 287-335.
- 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.
- Hayes, Bruce and Colin Wilson (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379-440.
- Hayes, Bruce, Kie Zuraw, Peter Siptár and Zsuzsa Londe (in press) Natural and unnatural constraints in Hungarian vowel harmony. To appear in *Language*.
- Jespersen, Otto (1901, English translation 1933). Notes on meter. In *Linguistica*, 249-274. Copenhagen: Levin and Munksgaard.
- Kiparsky, Paul (1975) Stress, syntax, and meter. *Language* 51, 576-616.
- Kiparsky, Paul (1977) "The rhythmic structure of English verse. *Linguistic Inquiry* 8, 189-248.
- Magnuson, Karl, and Frank G. Ryder. 1970. The study of English prosody: an alternative proposal," *College English* 31. 789-820.
- Selkirk, Elizabeth O. 1982. *Sound and Syntax: the Relation between Sound and Structure*, MIT Press.

- Selkirk, Elizabeth O. 1986. On derived domains in sentence phonology. *Phonology Yearbook* 3:371–405.
- Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30:945–982
- Youmans, Gilbert. 1989. Milton's meter. In Paul Kiparsky and Gilbert Youmans, eds., *Rhythm and Meter*, Academic Press, Orlando, FL.