LSA.365 | Variation in Optimality Theory, Handout #4
Arto Anttila

GRADIENT PHONOTACTICS IN OT

Arto Anttila
Stanford University

1. Introduction

(1) PHONOTACTICS is the study of permissible and impermissible phoneme combinations in a language.

(2) Phonotactic principles are known to be GRADIENT: lexical items can be more or less well-formed depending on the phoneme combinations they contain.

(3) Phonotatic gradience emerges in at least two ways:
  (a) Statistical overrepresentation/underrepresentation in the lexicon
  (b) "Wug words", e.g. *stin* > *smy* > *bzarshk*

(4) Literature on the former: Albright 2006; Greenberg and Jenkins 1964; Ohala and Ohala 1986; Coleman and Pierrehumbert 1997; Vitevitch, Luce, Charles-Luce, and Kemmerer 1997; Frisch, Large, and Pisoni 2000; Frisch and Zawaydeh 2001; and Bailey and Hahn 2001.

(5) Literature on the latter: Greenberg 1950; McCarthy 1988, 1994; Pierrehumbert 1993; Frisch, Pierrehumbert, and Broe 2004; Pater and Coetzee 2005 (Arabic); Berkley 1994a, 1994b, 2000; Coleman and Pierrehumbert 1997; Hammond 2004; Hay, Pierrehumbert, and Beckman 2004 (English); Pater and Coetzee 2005; Coetzee and Pater 2006 (Muna).

(6) Possible explanations:
  (a) GRAMMAR: abstract generalizations over natural classes
  (b) LEXICON: new words are supported by existing words
  (c) Both?

(7) Intuitively, the relative well-formedness of phonotactic combinations depends on markedness: marked combinations are less well-formed.

(8) The Complexity Hypothesis: The probability of an <input, output> mapping is inversely correlated with its grammatical complexity (to be defined shortly).

(9) Consequence: Much of gradient phonotactics is ranking-independent, hence universal, and thus does not have to be learned, cf. Hayes and Wilson 2006.

2. The Arabic example

(10) In Arabic, root morphemes normally consist of three consonants, e.g. *ktb* 'write'.

(11) A dissimilatory constraint against homorganic consonants in adjacent positions within the verbal root (Frisch, Pierrehumbert, and Broe 2004; Greenberg 1950; McCarthy 1988, 1994; Pater and Coetzee 2005; Pierrehumbert 1993).

(12) Thus, e.g. **fbm, *bfk, *kbm* are ill-formed (McCarthy 1988:88).

(13) The pattern is gradient: the more similar the consonants, the less frequently they co-occur in actual lexical items (Frisch, Pierrehumbert and Broe 2004).

(14) OBSERVED/EXPECTED (O/E) VALUE: The ratio of the observed number of occurring consonant pairs (O) to the number that would be expected if the consonants combined at random (E).

(15) Example (see e.g. Frisch, Pierrehumbert and Broe 2004:185)
- What is the O/E value for the sequence C1-C2?
- O = the number of times C1-C2 actually occurs in the dictionary.
- E = the number of times C1-C2 would occur if C1 and C2 combined freely.

(16) Calculate E as follows:
- Calculate the probability of C1 in the first position
- Calculate the probability of C2 in the second position.
- The probability of the pair is the product of these two probabilities.
- The expected frequency E is the product of the pair probability and the total number of consonant pairs.

(17) An O/E value > 1 indicates that there are more observed combinations than expected, i.e. the combination is favored.

(18) An O/E value < 1 indicates that there are fewer observed combinations than expected, i.e. the combination is disfavored.

(19) O/E values for pairs of adjacent consonants in Arabic verbal roots (Frisch, Pierrehumbert, and Broe 2004:186). The data are based on 2,674 Arabic roots taken from a dictionary of standard Arabic (Cowan 1979).

	labial	dorsal	coronal sonorant	coronal fricative	coronal plosive
labial	0.00				
dorsal	1.15	0.02			
coronal sonorant	1.18	1.48	0.06		
coronal fricative	1.31	1.16	1.21	0.04	
coronal plosive	1.37	0.80	1.23	0.52	0.14

labials:	b,f,m
dorsals:	k, g, q
coronal sonorants:	l, r, n
coronal fricatives:	θ, ð, s, z, $s^{ʕ}$, $z^{ʕ}$, ʃ
coronal plosives:	t, d, $t^{ʕ}$, $d^{ʕ}$

(20) The quantitative patterning of adjacent coronals
- (a) If the coronals are both sonorants, fricatives, or plosives, O/E is low;
- (b) If the coronals are fricative + plosive, O/E is higher;
- (c) If the coronals are sonorant + fricative or plosive, O/E is high.

(21) How to derive this pattern from the grammar?

3. Gradient phonotactics and T-orders

(22) Constraints for coronals (Pater &Coetzee 2005, see also McCarthy 1988, 1994, Padgett 1995)

FAITH	Input and output are identical.
OCP-COR	No adjacent coronals (e.g. /t-n/)
OCP-COR[−son]	No adjacent coronal obstruents (e.g. /t-s/)
OCP-COR[+son]	No adjacent coronal sonorants (e.g. /l-n/)
OCP-COR[−son][αcont]	No adjacent coronal obstruents agreeing in continuancy (e.g. /t-d/)

(23) The constraint violations for coronals

		FAITH	OCP-COR [−son, αcont]	OCP-COR [+son]	OCP-COR [−son]	OCP-COR
t-d	t-d		*		*	*
	OTHER	*				
t-s	t-s				*	*
	OTHER	*				
t-n	t-n					*
	OTHER	*				
l-n	l-n			*		*
	OTHER	*				

(24) The factorial typology for coronals

```
            #1    #2    #3    #4    #5    #6    #7
/t-d/:      t-d   t-d   OTHER OTHER OTHER OTHER OTHER
/t-s/:      t-s   t-s   t-s   t-s   OTHER OTHER OTHER
/t-n/:      t-n   t-n   t-n   t-n   t-n   t-n   OTHER
/l-n/:      l-n   OTHER l-n   OTHER l-n   OTHER OTHER
```

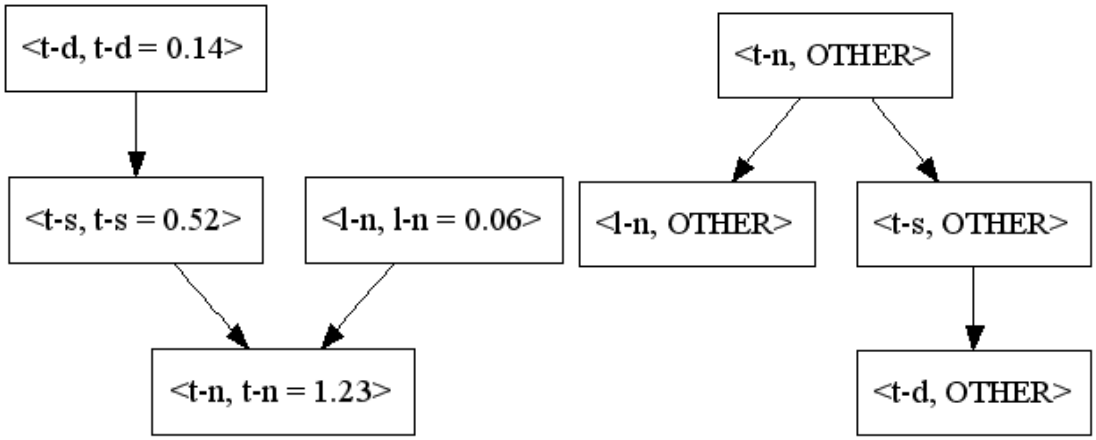
(25) An implicational universal

`<t-d, t-d> --> <t-s, t-s>` If /t-d/ is realized faithfully, so is /t-s/.

(26) T-order as pairs of <input, output> pairs

```
(a)   <t-d, t-d> --> <t-s, t-s>
(b)   <t-d, t-d> --> <t-n, t-n>
(c)   <t-s, t-s> --> <t-n, t-n>
(d)   <l-n, l-n> --> <t-n, t-n>
(e)   <t-s, OTHER> --> <t-d, OTHER>
(f)   <t-n, OTHER> --> <t-d, OTHER>
(g)   <t-n, OTHER> --> <t-s, OTHER>
(h)   <t-n, OTHER> --> <l-n, OTHER>
```

(27) T-order as a directed graph



(28) Observation: Implicational universals are reflected quantitatively in the Arabic lexicon. What is typologically marked is quantitatively underrepresented. Why?

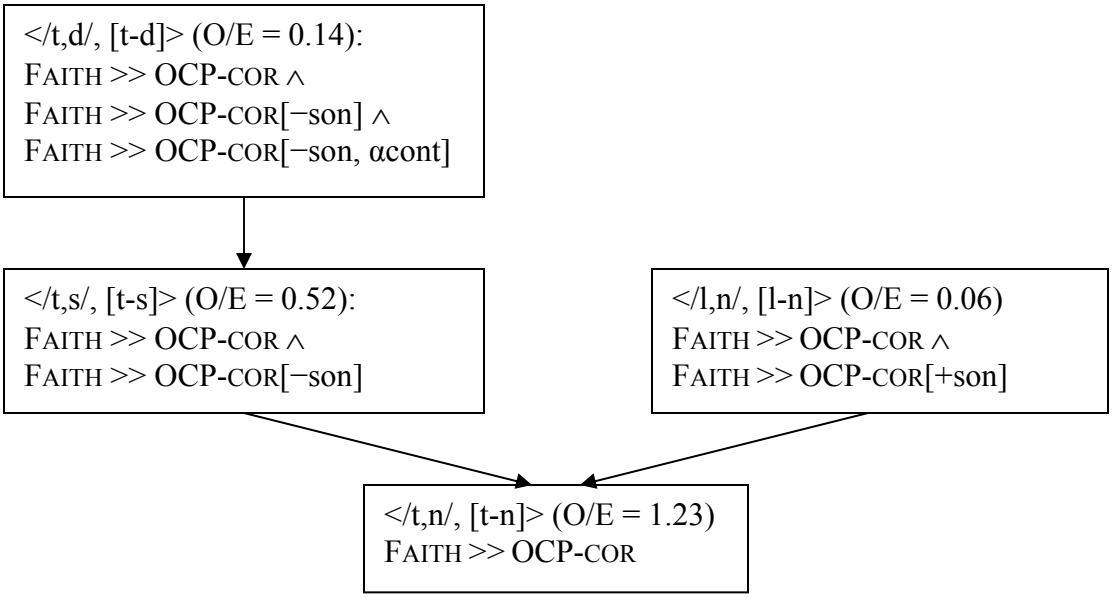
(29) Consider again the violation patterns for coronals:

		FAITH	OCP-COR [−son, αcont]	OCP-COR [+son]	OCP-COR [−son]	OCP-COR
t-d	t-d		*		*	*
	OTHER	*				
t-s	t-s				*	*
	OTHER	*				
t-n	t-n					*
	OTHER	*				
l-n	l-n			*		*
	OTHER	*				

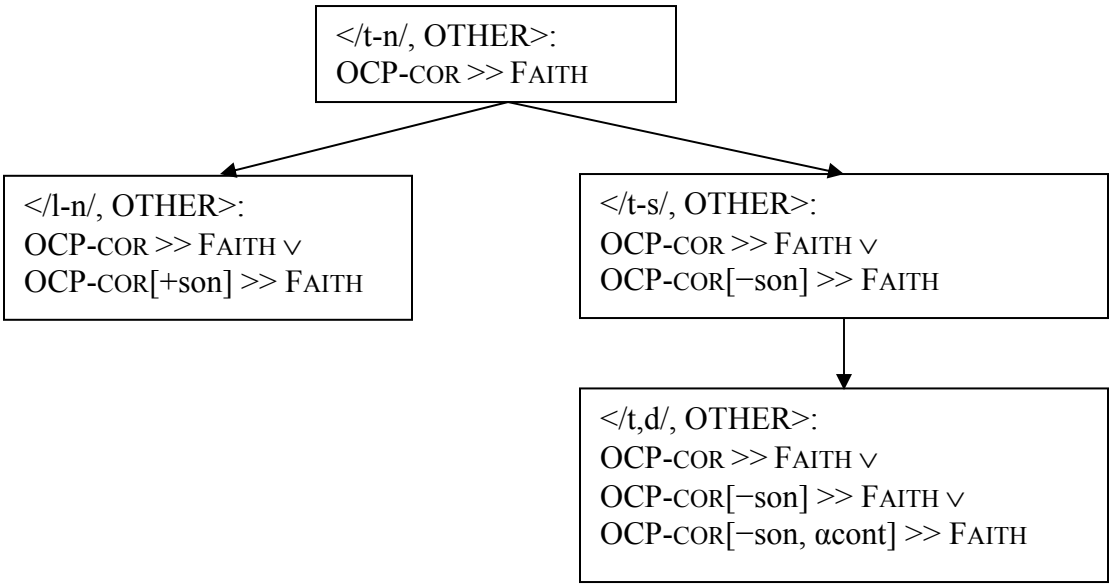
(30) The rankings required for the faithful mapping </t,d/, [t-d]>

FAITH >> OCP-COR ∧
FAITH >> OCP-COR[−son] ∧
FAITH >> OCP-COR[−son, αcont]

(31) Faithful mappings ordered by the amount of ranking information



(32) Unfaithful mappings ordered by the amount of ranking information



(33) Two different ways of looking at T-orders:
- (a) EXTENSIONALLY in terms of factorial typologies
- (b) INTENSIONALLY in terms of rankings

(34) Summary: T-orders arrange the <input, output> mappings by decreasing complexity, where complexity is measured in terms of the amount of ranking information.

(35) The Complexity Hypothesis (preliminary version): The O/E value of an <input, output> mapping is inversely correlated with its grammatical complexity.

4. Why should the Complexity Hypothesis be true?

(36) [Thanks to Michael Kenstowicz (p.c.) for raising these issues.]

(37) A language's vocabulary is largely a matter of history and the result of extragrammatical events such as language contact.

(38) Thus, it is probably not literally true that "when new words are added to the lexicon, forms that are more well-formed are more likely to be added than forms that are less well-formed" (Coetzee 2006a:382).

(39) What may be true is that over time, all else being equal, a language's lexical stock will favor less complex items over more complex items.

(40) For example, stems may form prosodically natural classes (monosyllabic, disyllabic, etc.), affixes may be restricted to natural classes (dentals, etc.) In this sense, the lexicon is a single system where *tout se tient*.

(41) How should all this be understood in terms of synchronic grammar?

(42) The answer is not obvious in OT. The Arabic ranking must be as in (43), since all these pairs are attested, but there is no sense in which the mapping <t-d, t-d> is "less optimal" than <t-n, t-n>.

(43) The ranking for Arabic (first attempt)

		FAITH	OCP-COR [−son, αcont]	OCP-COR [+son]	OCP-COR [−son]	OCP-COR
t-d	☞ t-d		*		*	*
	OTHER	*				
t-s	☞ t-s				*	*
	OTHER	*				
t-n	☞ t-n					*
	OTHER	*				
l-n	☞ l-n			*		*
	OTHER	*				

(44) An alternative view: The lexicon of a language is not phonologically homogeneous, but consists of subsystems, e.g. noun and verb phonologies (see e.g. Firth 1958, and more recently e.g. Itô and Mester 1995, 2002).

(45) Within OT, this view has been adopted by proponents of COPHONOLOGIES: (classes of) lexical items may differ in their constraint rankings (see e.g. Anttila 2002, 2006; Inkelas 1998; Inkelas and Zoll 2003; Orgun 1996; Raffelsiefen 1999; Zamma 2005).

(46) Proposal: Language users value simple lexical items over complex ones: a lexical item that requires more ranking information is at a disadvantage in comparison to a lexical item that requires less ranking information.

(47) It is this SIMPLICITY PRINCIPLE that is reflected quantitatively in the Arabic lexicon: lexical items that require more ranking information are less frequent than lexical items that require less ranking information.

5. Extending the analysis

(48) Check whether the analysis still works if we include all the distinct consonant pairs listed by Frisch, Pierrehumbert, and Broe (2004).

(49) The adjacent consonant pairs (Frisch, Pierrehumbert, and Broe 2004, 186)

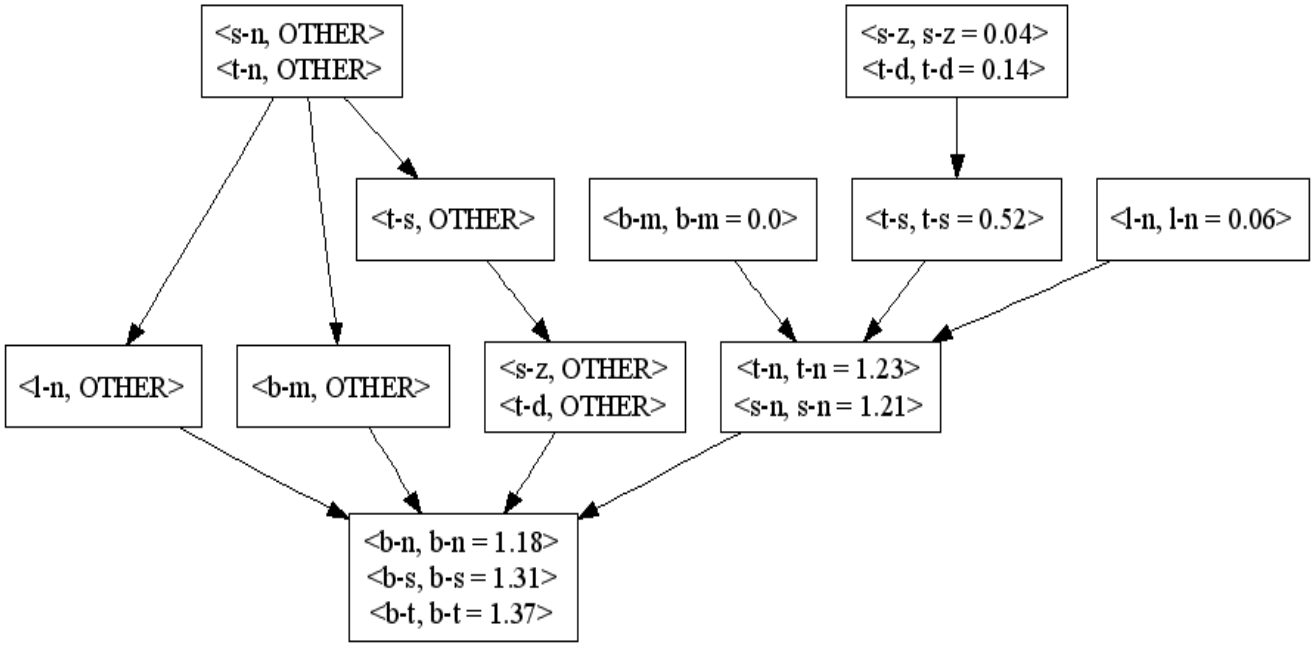
CONSONANT PAIR		EXAMPLE	O/E VALUE
LAB	LAB	b-m	0.00
LAB	COR[−son, −cont]	b-t	1.37
LAB	COR[−son, +cont]	b-s	1.31
LAB	COR[+son]	b-n	1.18
COR[−son, −cont]	COR[−son, −cont]	t-d	0.14
COR[−son, −cont]	COR[−son, +cont]	t-s	0.52
COR[−son, −cont]	COR[+son]	t-n	1.23
COR[−son, +cont]	COR[−son, +cont]	s-z	0.04
COR[−son, +cont]	COR[+son]	s-n	1.21
COR[+son]	COR[+son]	l-n	0.06

(50) Assumption: OCP-PLACE is STRINGENT: there is no markedness constraint that only targets coronals (OCP-COR), but there is a markedness constraint that targets all places (OCP-PLACE or OCP-LAB/COR) (Kiparsky 1994, De Lacy 2002).

(51) The complete violation tableau

		FAITH	OCP-LAB	OCP-LAB/COR	OCP-COR [+son]	OCP-COR [−son]	OCP-COR [−son, αcont]
b-s	b-s (1.31)						
	OTHER	*					
b-m	b-m (0.00)		*	*			
	OTHER	*					
b-n	b-n (1.18)						
	OTHER	*					
b-t	b-t (1.37)						
	OTHER	*					
t-s	t-s (0.52)			*		*	
	OTHER	*					
t-n	t-n (1.23)			*			
	OTHER	*					
s-n	s-n (1.21)			*			
	OTHER	*					
t-d	t-d (0.14)			*		*	*
	OTHER	*					
l-n	l-n (0.06)			*	*		
	OTHER	*					
s-z	s-z (0.04)			*		*	*
	OTHER	*					

(52) T-order for Arabic: precision = 0.837, recall: 0.8



(53) Caveat: O-values are typically rather low, precisely because of the strong OCP effect. This makes it hard to test statistical significance.

(54) Frisch et al. (2004) compared models using the residual sum of squares between the data and the model predictions and compared models in terms of better/worse fit (Frisch, p.c.).

(55) That said, the OT model derives explicit and fine-grained predictions that can be in principle tested on experimental data.

6. Summary

(56) T-orders are a consequence of standard OT, not a new theoretical device.

(57) A different issue: How should T-orders be interpreted empirically?

(58) Our proposal is the Complexity Hypothesis. This is an empirical claim that may be true or false.

(59) The Complexity Hypothesis makes specific and often surprising predictions about phonotactic patterns, given a set of constraints.

(60) For example, Pater and Coetzee 2005's constraints predict the following:
(a) The orderings [t-d] ≤ [t-s] ≤ [t-n] and [l-n] ≤ [t-n] should be universal.
(b) The orderings between [l-n] and [t-s] and between [l-n] and [t-d] should vary from language to language.

(61) Q: Why do gradient phonotactic generalizations exist?
A: Because some phoneme combinations are more marked than others.

(62) Q: How is this formalized in the grammar?
A: Marked phoneme combinations are more complex than unmarked ones.

(63) We have formalized these intuitions in terms of T-orders.

References

Albright, Adam. 2006. 'Gradient phonotactic effects: lexical? grammatical? both? neither?' Paper presented at the 80th Annual Meeting of the Linguistic Society of America, Albuquerque, NM.

Anttila, Arto. 2002. 'Morphologically Conditioned Phonological Alternations', *Natural Language and Linguistic Theory* 20, 1-42. Also on http://roa.rutgers.edu/.

Anttila, Arto. 2006. 'Variation and opacity', *Natural Language and Linguistic Theory* 24, 893-944. Also on http://roa.rutgers.edu/.

Anttila, Arto. 2007a. 'Word stress in Finnish', Paper presented at the 81st Annual Meeting of the Linguistic Society of America, Anaheim, CA. Handout available at http://www.stanford.edu/~anttila/research/papers.

Anttila, Arto. 2007b. 'Prosodic constraints on constituent ordering', Paper to be presented at the 26th West Coast Conference on Formal Linguistics, University of California, Berkeley.

Anttila, Arto and Curtis Andrus. 2006. 'T-orders', Ms. (comes with software), available at http://www.stanford.edu/~anttila/research/torders/t-order-manual.pdf.

Anttila, Arto, Vivienne Fong, Stefan Benus, and Jennifer Nycz. in progress. 'Consonant clusters in Singapore English', Ms., Stanford University, Columbia University, and New York University.

Bailey, Todd M. and Ulrike Hahn. 2001. 'Determinants of wordlikeness: Phonotactics or lexical neighborhoods?' *Journal of Memory and Language* 44, 568-591.

van den Berg, René. 1989. *A Grammar of the Muna Language*, Foris, Dordrecht.

van den Berg, René and La Ode Sidu. 1996. *Muna-English Dictionary*, KITLV Press, the Netherlands.

Berkley, Deborah Milam. 1994a. 'Variability and Obligatory Contour Principle Effects', in Beals et al., (eds.), *CLS 30: Proceedings of the 30th Annual Meeting of the Chicago Linguistic Society, Volume 2: The Parasession on Variation and Linguistic Theory,* Chicago Linguistic Society, Chicago. pp. 1-12.

Berkley, Deborah Milam. 1994b. 'The OCP and gradient data', *Studies in the Linguistic Sciences* 24, 59-72.

Berkley, Deborah Milam. 2000. *Gradient Obligatory Contour Principle Effects*, Ph.D. dissertation, Northwestern University.

Blumenfeld, Lev. 2005. 'Matching ictus and stress in Latin hexameter endings', Paper presented at the 80th Annual Meeting of the Linguistic Society of America, Albuquerque, NM.

Coetzee, Andries W. 'Variation as accessing "non-optimal" candidates', *Phonology* 23(3), 337-385.

Coetzee, Andries W. and Joe Pater. 2006. 'Lexically ranked OCP-Place constraints in Muna', Ms., University of Michigan and University of Massachusetts, Amherst. ROA-842.

Coleman, John and Janet Pierrehumbert. 1997. 'Stochastic phonological grammars and acceptability', in *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*, Association for Computational Linguistics, Somerset. pp. 49-56.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. 'Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords', *Journal of Memory and Language* 42, 481-496.
Frisch, Stefan and Bushra Zawaydeh. 2001. 'The Psychological Reality of OCP-Place in Arabic', *Language* 77, 91-106.
Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. 'Similarity avoidance and the OCP', *Natural Language and Linguistic Theory* 22(1), 179-228.
Greenberg, Joseph. 1950. 'The patterning of root morphemes in Semitic', *Word* 6, 162-181.
Greenberg, J. H. and J. J. Jenkins. 1964. 'Studies in the psychological correlates of the sound system of American English', *Word* 20, 157-177.
Hammond, Michael. 2004. 'Gradience, phonotactics, and the lexicon in English phonology', *International Journal of English Studies*, 4(2), 1-24.
Hay, Jennifer, Janet Pierrehumbert, and Mary Beckman. 2004. 'Speech Perception, Well-Formedness, and the Statistics of the Lexicon', in J. Local, R. Ogden, and R. Temple, (eds.), *Papers in Laboratory Phonology VI*, Cambridge University Press, Cambridge UK, pp. 58-74. Also at http://www.ling.canterbury.ac.nz/people/hay.shtml.
Hayes, Bruce, Bruce Tesar, and Kie Zuraw. 2003. 'OTSoft 2.1, software package', available at http://www.linguistics.ucla.edu/people/hayes/otsoft/.
Inkelas, Sharon. 1998. 'The theoretical status of morphologically conditioned phonology: a case study of dominance effects', in Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1997*, Kluwer, Dordrecht, pp. 121-155.
Inkelas, Sharon and Cheryl Zoll. 2003. 'Is grammar dependence real?', ROA-587.
Kruskal, J.B. 1983. 'An overview of sequence comparison: Time warps, string edits, and macromolecules', *SIAM Review* 25(2), 201-237.
Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts.
McCarthy, John. 1988. 'Feature geometry and dependency: A review', *Phonetica* 45, 84-108.
McCarthy, John. 1994. 'The phonetics and phonology of Semitic pharyngeals', in Patricia Keating, (ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology* 3, Cambridge University Press, Cambridge, pp. 191-233.
Ohala, J. and M. Ohala. 1986. 'Testing hypotheses regarding the psychological manifestation of morpheme structure constraints', in *Experimental Psychology*, Academic Press, Orlando, FL, pp. 239-252.
van Oostendorp, Marc, and Frans Hinskens. 2007. 'Segmental inventories and T-orders: Evidence from Dutch dialects', Ms. Meertens Institute.
Orgun, Cemil Orhan. 1996. *Sign-Based Morphology and Phonology with special attention to Optimality Theory*, Ph.D. dissertation, Department of Linguistics, University of California, Berkeley, ROA-171.
Padgett, Jaye. 1995. *Stricture in Feature Geometry*. Dissertations in Linguistics, Center for the Study of Language and Information, Stanford, CA.

Pater, Joe and Andries Coetzee. 2005. 'Lexically Specific Constraints: Gradience, Learnability, and Perception', *Proceedings of the Phonology-Morphology Circle of Korea,* Seoul, Korea, pp. 85-119.

Pierrehumbert, Janet. 1993. 'Dissimilarity in the Arabic verbal roots', in A. Schafer, (ed.), *NELS 23: Proceedings of the North East Linguistic Society,* GLSA, University of Massachusetts, Amherst, pp. 367-381.

Prince, Alan. 2002a. 'Entailed Ranking Arguments', ROA-500.

Prince, Alan. 2002b. 'Arguing Optimality', ROA-562.

Prince, Alan. 2006a. 'Implication & Impossibility in Grammatical Systems: What it is & How to find it', ROA-880.

Prince, Alan. 2006b. 'No More than Necessary: beyond the 'four rules', and a bug report', ROA-882.

Prince, Alan. 2007. 'The pursuit of theory', in Paul de Lacy (ed.), *The Cambridge Handbook of Phonology*, Cambridge University Press, Cambridge. pp. 33-60.

Prince, Alan and Paul Smolensky 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*, Blackwell Publishing, Malden, Massachusetts.

Raffelsiefen, Renate. 1999. 'Phonological constraints on English word formation', in Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1998*, pp. 225-287.

Vitevitch, M. S., P. A. Luce, J. Charles-Luce, and D. Kemmerer. 1997. 'Phonotactics and syllable stress: Implications for the processing of spoken nonsense words', *Language and Speech* 40, 47-62.

Zamma, Hideki. 2005. 'Predicting Varieties: Partial Orderings in English Stress Assignment', ROA-712.