# Investigating phonological abstraction through feature induction

Features in Phonology, Morphology, Syntax: What are they?
Universitetet i Tromsø, October 31 2013

Aleksei Nazarov, University of Massachusetts at Amherst

anazarov@linguist.umass.edu

#### Overview

- Introduction
  - should grammars always refer to features?
  - approach from perspective of machine learning

#### Overview

- Introduction
  - should grammars always refer to features?
  - approach from perspective of machine learning
- Computational simulation: how does a learner abstract over domains of application?
  - model, data, method
  - results: grammars with features <u>in some</u> <u>constraints only</u>

#### Overview

- Introduction
  - should grammars always refer to features?
  - approach from perspective of machine learning
- Computational simulation: how does a learner abstract over domains of application?
  - model, data, method
  - results: grammars with features <u>in some</u> <u>constraints only</u>
- Discussion: implications of grammars referring to features as well as other units

# Introduction: background

- Features help generalize over domains of application of rules or constraints
- Phonology: features generalize over segment/ phoneme categories

E.g., 
$$/-z/ \rightarrow [-s] / [p,t,k,f,\theta,s,\int,tf]_ \Rightarrow$$
  
 $/-z/ \rightarrow [-s] / [-voice]_$ 

# Introduction: background

• Question:

Is it always advantageous (both for the analyst and the speaker) to state every rule or constraint in the grammar in terms of features?

In other words: is it unreasonable for grammar to refer to sound event through levels of abstraction other than features?

(Not counting prosodic units, suprasegmentals)

# Introduction: background

Phonology: canonical answer is "yes"

## Introduction: background

- Phonology: canonical answer is "yes"
- Chomsky & Halle (1968):
  - adapting categorical versions of phonetic features is most economical hypothesis of representation

## Introduction: background

- Phonology: canonical answer is "yes"
- Chomsky & Halle (1968):
  - adapting categorical versions of phonetic features is most economical hypothesis of representation
  - establishes preference for phonetically natural rules

(see Chomsky & Halle 1968, Postal 1968, Kenstowicz & Kisseberth 1979 for more)

## Introduction: background

- Phonology: canonical answer is "yes"
- Chomsky & Halle (1968):
  - adapting categorical versions of phonetic features is most economical hypothesis of representation
  - establishes preference for phonetically natural rules

(see Chomsky & Halle 1968, Postal 1968, Kenstowicz & Kisseberth 1979 for more)

 Models with richer representations lead to longer grammars, therefore are disfavored

# Introduction: empirical issue

Phonological patterns may apply to groups of segments, or to single segments.

- Phonological patterns may apply to groups of segments, or to single segments.
- English (Jensen 1993, Mielke 2007):
  - sibilants [s,z,∫,ʒ,tʃ,ʤ] may not precede [s,z] word-finally: \*[bʌs-s, bʌz-z, pætʃ-s, peɪʤ-z]

```
p t k Red: disallowed before [s,z] word-finally b d g f θ s ∫ tf v ð z 3 d₃ m n ŋ w ɹ l j
```

- Phonological patterns may apply to groups of segments, or to single segments.
- English (Jensen 1993, Mielke 2007):
  - only [s] may start a three-consonant wordinitial cluster: [strit], \*[ftrit, ntrit, tftrit]

```
p t k Red: disallowed before [s,z] word-finally
b d g Purple: allowed as C1 in word-initial CCC
f θ s ∫ tf
v ð z 3 d3
m n n
w μ i
```

- Phonological patterns may apply to groups of segments, or to single segments.
  - P-base cross-linguistic database of phonological classes (Mielke 2007):
    - 13 patterns encoded as applying to one segment
    - 11 additional cases (apply to all segments but one) found by manual search of languages starting with A alone

# Introduction: empirical issue

 One-segment classes may be represented as intersections of a number of features

- One-segment classes may be represented as intersections of a number of features
  - -e.g., [s] is equivalent to [+ant,-voice,+strid]

- One-segment classes may be represented as intersections of a number of features
  - -e.g., [s] is equivalent to [+ant,-voice,+strid]

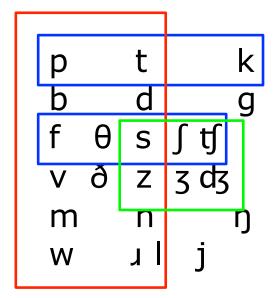
```
p t k Red: [+anterior]
b d g
f θ s ∫ t∫
v ð z 3 d
m n ŋ
w ɹ l j
```

- One-segment classes may be represented as intersections of a number of features
  - -e.g., [s] is equivalent to [+ant,-voice,+strid]

```
p t k Red: [+anterior]
b d g Blue: [-voice]
f θ s ∫ tf
v ð z 3 d3
m n ŋ
w ɹ l j
```

# Introduction: empirical issue

- One-segment classes may be represented as intersections of a number of features
  - -e.g., [s] is equivalent to [+ant,-voice,+strid]



Red: [+anterior]

Blue: [-voice]

Green: [+strident]

## Introduction: always features?

- Featural representation of one-segment class will always be longer and more complex
- Is it desirable (for analyst/speaker) to represent one-segment classes in this way?

## Introduction: always features?

- Featural representation of one-segment class will always be longer and more complex
- Is it desirable (for analyst/speaker) to represent one-segment classes in this way?
  - If features are *a priori* specified as building blocks of grammars: yes
  - Is this still the case when this *a priori* assumption is taken away?

# Introduction: machine learning

- I will approach this question in terms of machine learning
- Given a choice between representing a pattern in terms of segments and in terms of features:
  - How will data containing both one-segment and multi-segment patterns be learned?
  - Learning algorithm not explicitly instructed to aim for a certain level of abstraction

# Introduction: machine learning

- Possible outcomes:
  - 1.The grammars have constraints referring only to segments
  - 2. The grammars have constraints referring only to features
  - 3. The grammars have constraints referring to both features and segments

## Introduction: assumptions

- Essential assumptions for this simulation:
  - 1. Atomic segment units are available to the language user:
    - active in on-line processing of speech (Jesse et al. 2007, Nielsen 2011)
    - active in phonological processes, e.g., consonant OCP

(Coetzee & Pater 2008 and references therein)

## Introduction: assumptions

- Essential assumptions for this simulation:
  - 2. Phonological features are learned:
    - assuming universal features, the same feature is realized differently across languages

(Cho & Ladefoged 1999)

 therefore, phonetic information cannot be sufficient for mapping perception/articulation to features

## Introduction: assumptions

- Essential assumptions for this simulation:
  - 2. Phonological features are learned:
    - contextual information must be used
    - grammar contains contextual information
    - use contextual information from grammar (rather than contextual information outside of grammar)

(see Mielke (2004) on learning features from phonological patterns)

# Introduction: assumptions

- Consequences of these assumptions:
  - 1. Segment-to-feature mapping must be learned simultaneously with grammar
  - Constraints/rules referring to features gradually become available during grammar learning process

# Introduction: assumptions

- Non-essential working assumptions:
  - Features are induced only from contextual information: no phonetic content

(Substance-free phonology: Morén 2006, 2007 (and many others))

All phonological constraints are induced instead of innate

(see Hayes & Wilson 2008 on constraint induction)

## Introduction: summary

- Question: Is it always advantageous (both for the analyst and the speaker) to state every constraint in the grammar in terms of features?
- Crucial empirical phenomenon: one-segment patterns
- Learning one-segment and multi-segment patterns: all-feature grammars as outcome?
- Preview: segment/feature grammars obtained

#### Simulation: overview

- Machine learning simulation based on paradigm established by Hayes & Wilson (2008):
  - phonotactic constraint-based grammar is built up from positive data
  - violable constraints selected and weighted to optimally predict the attested data

#### Simulation: overview

- Departure from Hayes & Wilson's learner:
  - features are not built into the model, but induced at intermediate stages of grammar learning
- Questions:
  - will features be learned at all?
  - will all constraints in grammars learned by this procedure always use features?

#### Simulation: model

- Maximum Entropy model
   (Della Pietra et al. 1997, Hayes & Wilson 2008)
  - probability distribution over possible representations based on weighted violable constraints (à la OT/Harmonic Grammar)
  - constraints weighted to make this distribution maximally similar to what is observed

(see Appendix for more)

#### Simulation: model

- Regularization:
  - Optimization of constraint weights constrained by L2 prior (Hastie et al. 2009):
    - keeps sum of constraint weights as small as possible
    - encourages more general constraints:
       one general constraint with larger weight
       yields smaller sum of weights
       than several specific constraints with smaller
       weights

#### Simulation: model

- Information gain:
  - Value which estimates how much a constraint will improve the <u>current grammar</u> (bring it closer to predicting the observed data)
  - Information gain of a constraint correlates with how accurately it captures a (sub)pattern in the data

(see Appendix for more)

#### Simulation: model

- Constraints:
  - phonotactic constraints against two- and threeelement sequences of word-boundaries, segments or features
  - examples: \*#m, \*km, \*u[labial]u

#### Simulation: model

- Constraints:
  - selected probabilistically based on information gain:
    - start with random seed constraint

```
(subject to information gain threshold)

e.g. *#pi
```

 seed constraint repeatedly manipulated until this does not lead to increase in information gain

- Features found by clustering information gain of closely related constraints
  - Intuition:
     a feature denotes a class of segments that
     participates in the same phonological pattern

- Features found by clustering information gain of closely related constraints
  - Implementation:

     a feature denotes a class of segments which yields high-valued constraints when inserted in the same context

	i	a	u	р	t	k	b	d	g	m	n	α
*#_	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.015	0.015	0.015

- Features found by clustering information gain of closely related constraints
  - Cluster analysis (Mixture of Gaussians, Everitt 2011)
     divides same-context constraints into high and low information gain value clusters (whenever appropriate)

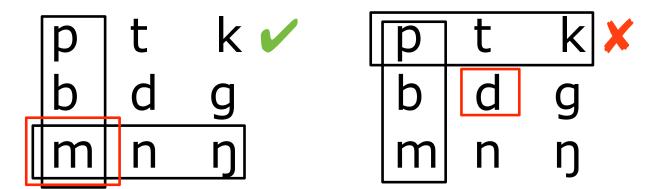
	i	a	u	p	t	k	b	d	g	m	n	α
*#	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.0020	0.015	0.015	0.015

- Features found by clustering information gain of closely related constraints
  - Focus segments extracted from cluster of high information-value constraints
  - Feature label assigned to these segments (phonetics not taken into account labels are arbitrary) [nasal]

	i	а	u	p	t	k	b	d	g (	m/	n	Q
*#_	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.015	0.015	0.015

#### Simulation: data

- Nature of data to consider:
  - both one-segment and multi-segment patterns must be present
  - single segment in one-segment pattern must be representable as intersection of segment classes appealed to in multi-segment patterns



#### Simulation: data

Example: English (Jensen 1993, Mielke 2007)

#### Simulation: data

Example: English (Jensen 1993, Mielke 2007)

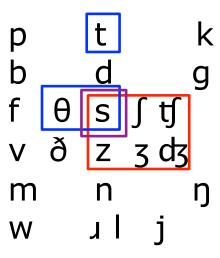
```
p t k
b d g
f θ s ∫ tf
v ð z z d
m n n
w ɹ l j
```

Red: disallowed before [s,z] word-finally

Blue: allowed as C3 in word-final CCC

#### Simulation: data

Example: English (Jensen 1993, Mielke 2007)



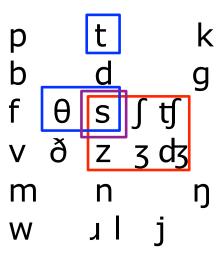
Red: disallowed before [s,z] word-finally

Blue: allowed as C3 in word-final CCC

Purple: allowed as C1 in word-initial CCC

#### Simulation: data

Example: English (Jensen 1993, Mielke 2007)



Red: disallowed before [s,z] word-finally

Blue: allowed as C3 in word-final CCC

Purple: allowed as C1 in word-initial C

 Other examples like this found in, e.g., Yoruba (Pulleyblank 1988)

#### Simulation: data

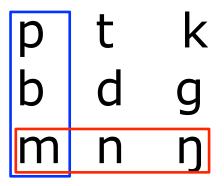
The actual data used for the simulations was a toy language which shared the crucial properties of these examples:

```
p t k
b d g
m n n
```

Red: no nasals word-initially

#### Simulation: data

The actual data used for the simulations was a toy language which shared the crucial properties of these examples:

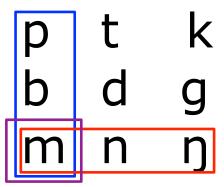


Red: no nasals word-initially

Blue: no labials between high vowels [i,u]

#### Simulation: data

The actual data used for the simulations was a toy language which shared the crucial properties of these examples:



Red: no nasals word-initially

Blue: no labials between high vowels [i,u]

Purple: no [m] word-finally

#### Simulation: data

The actual data used for the simulations was a toy language which shared the crucial properties of these examples:

 All possible CVCVC forms obeying these restrictions present in input to the learner

# Simulation: procedure

- Initial state: no constraints, features unavailable
- All potential representations (given in segments) equally probable

# Simulation: procedure

- Initial state: no constraints, features unavailable
- All potential representations (given in segments) equally probable
- All CVCVC sequences over toy language inventory are potential representations
- Observed forms have no initial nasals, no labials between high Vs, no final [m]

```
possible: ... pada<u>m</u> padan ... <u>n</u>itun ditun d<u>ibu</u>n
```

observed: ... padan ... ditun

#### Simulation: method

 Step 1: Find a group of constraints which forms a local peak in gain value

These have higher information gain than, e.g., \*#p, \*am, \*n:

\*#p, \*am, \*n ban (more) observed forms in the data and bring the empty grammar less close to predicting the observed data

#### Simulation: method

 Step 2: Find all possible contexts that can be made from these constraints.

The constraints {\*#m,\*#n,\*#ŋ} can be factored into the following contexts

- \*#\_
- \* m
- \* n
- \*\_ŋ

#### Simulation: method

 Step 3: for every context, find if there is a cluster of segments which yields a high information gain value when inserted in that context; assign feature labels to those clusters

	i	a	u	р	t	k	b	d	g	m	n	n
*#_	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.015	0.015	0.015

$$[m, n, \eta] \Rightarrow [nasal]$$

#### Simulation: method

 Step 4: add the selected constraints to the grammar, and optimize their weights

#### Grammar:

- Steps 1-4 repeated until final goal is reached (observed data have at least 95% total likelihood)
- Features induced at step 3 available for use in constraints at next occurrence of step 1
  - Once \*#m, \*#n, \*#ŋ are in the grammar, and the feature label [nasal] = [m, n, ŋ] is induced,
  - the constraint \*#[nasal] becomes available

- E.g., \*#[nasal] has high information gain value (not in current grammar, tightly fits data pattern)
- If selected and weighted, \*#[nasal] takes away all the weight of \*#m, \*#n, \*#n
- zero weight equivalent to absence from grammar

```
*#[nasal]: 8

*#m: 6

*#n: 0

*#n: 0

*#n: 0

*#n: 0
```

- Reset to 0 because of regularization prior:
  - higher weight on one constraint is better than lower weights on three constraints combined
- This effect occurs when the candidates punished by a new constraint are a strict superset of those punished by individual existing constraints:
  - \*#[nasal] *versus* \*#m, \*#n, \*#ŋ
  - \*[hi][labial][hi] versus \*ibi, \*ibu, \*umi ...

- Reset to 0 does not happen when feature-based constraint and segment-based constraint are homonymous:
  - -\*[labial,nasal]# = \*m#

- Reset to 0 does not happen when feature-based constraint and segment-based constraint are homonymous:
  - -\*[labial,nasal]# = \*m#
- Homonymous feature-based constraint has lower information gain (repeats existing constraint)
   \*[lab,nas]# less likely to be selected
- Even when it is selected, no reset to 0 \*m# retains some weight next to \*[lab,nas]#

#### Simulation: results

- 31 out of 32 runs yielded grammars referring to both segments and features
- Most frequent grammar: \*#[nasal], \*[high][labial][high], \*m#
- One all-feature grammar:\*#[nasal], \*[hi][labial][hi], \*[labial,nasal]#
- All other grammars were variations of the most frequently observed grammar (see Appendix)

#### Simulation: results

- The learner strongly prefers a segmental representation for the one-segment pattern, and a featural representation for the multi-segment patterns.
- By extrapolation, languages with at least one one-segment pattern are expected not to represent that one-segment pattern (entirely) in terms of features.

#### Discussion

- Machine learning simulation shows:
  - when *a priori* assumption of all-feature grammars is lifted:
  - despite bias in favor of generalization,
  - one-segment patterns not represented in terms of features
- This is because features are more efficient <u>only</u> for multi-segment patterns

#### Discussion

- These results show that:
  - features can be learned in a bottom-up fashion from phonological patterns
  - grammars that represent one-segment patterns without features emerge despite bias towards generalization (from regularization)

#### Discussion

- These results show that:
  - features can be learned in a bottom-up fashion from phonological patterns (see also Archangeli et al. 2012)
  - grammars that represent one-segment patterns without features emerge despite bias towards generalization (from regularization)

(Procedure relies only on structural factors: these methods may also be applied to other domains of language, e.g., syntax)

- Implication for (phonological) analysis:
  - when a (phonological) pattern is analyzed, it is not trivial that it is stated in terms of features
  - rather, question of appropriate level of abstraction asked for every pattern

- Implication for (phonological) analysis:
  - when a (phonological) pattern is analyzed, it is not trivial that it is stated in terms of features
  - rather, question of appropriate level of abstraction asked for every pattern
- Why would level of abstraction matter?

- There are psycholinguistic techniques to probe into levels of abstraction:
  - Bach testing (Halle 1978)
  - Priming (Jesse et al. 2007)
  - Talker adaptation (McQueen et al. 2006, Nielsen 2011)
- Ergo: level of abstraction in hypothesized rules/ constraints matters empirically
- Important direction for future research

- Another consequence of grammars with both featural and lower-order descriptions:
  - same sound event may be described at different levels of abstraction
     e.g., [m] or [labial,nasal]
  - this means: multiple autonomous levels of representation for sounds

- This property is reminiscent of models such as
  - Turbidity (Goldrick 2001)
  - Abstract Declarative Phonology (Bye 2006)
  - Colored Containment (Van Oostendorp 2004, 2008)
  - Bidirectional Phonology (Boersma 2007)
- Grammars with multiple levels of abstraction need little extension to have the extra power of such models (Nazarov 2012, 2013)
- Another direction for further investigation

#### Conclusion

- Are features always better for representing phonological patterns?
- Investigation through machine learning of features:
  - no: one-segment patterns favor representation by segment units
- Grammars which refer both to features and lower-order units (segments) are worthy of consideration by speakers and analysts

# UMassAmherst Thank you! Aleksei Nazarov, University of Massachusetts at Amherst 72

## Acknowledgements

- Many thanks to:
- Kristine Yu
- Brian Dillon
- Tom Roeper
- Joe Pater
- John Kingston
- John McCarthy
- participants of the UMass Sound Seminar and the UMass Phonology Reading Group

#### References

**Archangeli**, D., J. Mielke & D. Pulleyblank. 2012. 'From Sequence Frequencies to Conditions in Bantu Vowel Harmony: Building a grammar from the ground up.' In: B. Botma & R. Noske (eds.), *Phonological Explorations: Empirical, Theoretical and Diachronic Issues*, Berlin: Mouton de Gruyter, pp. 191-222.

Boersma, P. 2007. Some listener-oriented accounts of h-aspiré in French. Lingua, 117, 1989-2054.

**Blaho**, S., P. Bye & M. Krämer (eds.). 2007. *Freedom of Analysis?* Berlin/New York: Mouton de Gruyter.

**Bye**, P. 2006. *Grade alternation in Inari Saami and Abstract Declarative Phonology*. Ms., Universitetet i Tromsø.

**Cho**, T. & P. Ladefoged. 1999. 'Variation and universals in VOT: evidence from 18 languages.' *Journal of Phonetics*, 27, 2, 207--229.

Chomsky, N. & M. Halle. 1968. The sound pattern of English. New York (NY): Harper and Row.

Coetzee, A. & J. Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory*, 26, 289-337.

**Della Pietra**, S., V.J. Della Pietra & J.D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 380–393.

#### References

Everitt, B. 2011. Cluster analysis. 5th edition. Chichester, West Sussex: Wiley.

**Goldrick**, M. 2001. Turbid output representations and the unity of opacity. In: M. Hirotani, A. Coetzee, N. Hall & J.-Y. Kim (eds.), *Proceedings of the Northeast Linguistic Society 30, Rutgers University*, Amherst, MA: GLSA, pp. 231-245.

**Halle**, M. 1978. Knowledge unlearned and untaught: what speakers know about the sounds of their language. In: M. Halle, J. Bresnan & G.A. Miller (eds.), *Linguistic theory and psychological reality*, Cambridge, MA and London: MIT Press, pp. 294-303.

**Hastie**, T, R. Tibshirani & J. Friedman. 2009. *The elements of statistical learning*. Second edition. New York: Springer.

**Hayes**, B. & C. Wilson. 2008. 'A maximum entropy model of phonotactics and phontactic learning.' *Linguistic Inquiry*, 39, 379-440.

**Jesse**, A., J.M. Page & M. Page (2007). 'The locus of talker-specific effects in spoken-word recognition'. In: Proceedings of ICPhS XVI, pp. 1921-1924. **Jensen**, J. 1993. *English phonology*. Amsterdam: John Benjamins.

**McQueen**, J.M., A. Cutler & D. Norris. 2006. Phonological abstraction in the mental lexicon. *Cognitive Science*, 30, 1113-1126.

#### References

**Mielke**, J. 2004. *The emergence of distinctive features*. Doctoral dissertation, Ohio State University.

Mielke, J. 2007. P-base, version 1.92. Software, University of Ottawa.

**Morén**, B. 2006. Consonant–vowel interactions in Serbian: features, representations and constraint interactions. *Lingua*, 116, 8, 1198–1244.

**Morén**, B. 2007. 'The division of labor between segment-internal structure and violable constraints'. In: Blaho, Bye & Krämer (2007), pp. 313–344.

**Nazarov**, A. 2012. *Phonological opacity as differential classification of sound events*. Ms., University of Massachusetts Amherst.

**Nazarov**, A. 2013. *Phonological opacity as differential classification of sound events*. Talk given at the University of Amsterdam on 1/10/2013.

**Oostendorp**, M. van. 2004. *The theory of faithfulness*. Ms., Meertens Instituut.

Oostendorp, M. van. 2008. Incomplete Devoicing in Formal Phonology. *Lingua*, 118, 1362-1374.

Pulleyblank, D. 1988. Vocalic underspecification in Yoruba. *Linguistic Inquiry*, 19, 2, 233-270.

# Appendix: Maximum Entropy model

Observed distribution p

$$p(x) = count(x) / \sum_{y \in \Omega} count(y)$$

 Predicted distribution q: based on harmony scores H for every candidate

$$H(x) = \Sigma (w_i \times C_i(x))$$

$$q(x) = e^{H(x)} / \sum_{y \in \Omega} e^{H(y)}$$

 $\Omega$  stands for the set of possible representations

# Appendix: Maximum Entropy model

 Objective of the model: manipulate weights to minimize K-L divergence of observed distribution from predicted distribution

$$D_{KL}(t || w) = \Sigma [t(x) * ln(t(x) / w(x))]$$

Obj = min 
$$[D_{KL}(p || q) + \sum_{w \in W} [(w - \mu)^2 / 2\sigma]]$$

regularization term;

$$\mu = 0 \text{ and } \sigma = 10,000$$

# Appendix: Information gain

- Let C\* be a proposed new constraint, and w\* its weight
- Let q' be the distribution predicted by the current grammar augmented with C\* with weight w\*
- Information gain: maximum descent in K-L divergence of observed from predicted when C\* is added to the grammar

(L2 regularization with  $\mu = 0$  and  $\sigma = 10,000$  added to this maximization also)

$$G(w^*,C^*) = \max_{w^*} [D_{KL}(p || q) - D_{KL}(p || q')]$$

## Appendix: Results

- Word-initial pattern:
  - 26 grammars: represented by \*#[nasal]
  - 3 grammars: \*#[nasal], \*#[nasal]V
  - 3 grammars:

(42) the three runs at which the word-initial restriction was represented by non-overlapping constraints

Run 11			Run 16		_	Run 17			
Constraint	Traditional notation	Weight	Constraint	Traditional notation	Weight	Constraint	Traditional notation	Weight	
*#m	*#m	2.68	*#{nŋ}	*#[nasal, -labial]	2.78	*#{nŋ}	*#[nasal, -labial]	3.37	
*#{nŋ}	*#[nasal, -labial]	1.12	*#{mŋ}	*#[nasal, -coronal]	2.78	*#m	*#m	2.68	
*#{nŋ} {aiu}	*#[nasal, -labial]V	1.12							
*#{nŋ}	*#[nasal,- labial]	1.12							

## Appendix: Results

- Word-medial pattern:
  - Combination of one or more of the following constraints:

(43) a survey of all 18 constraints attested in the final grammars which represented (part of) the word-medial pattern

```
*{iu}{pbm}{iu}
                             *{iu} {pbm}
                                                      *{aiu}m
              *{iu}{pbm}u
                                             *mi
*{pbm}{iu}
                                                      *m{aiu}
                                             *mu
                             *{iu}{pb}
*{iu}{pb}{iu}
              *u{pbm}{iu}
*{iu}b{iu}
               *u{pm} {iu}
                              *{iu}m
*{iu}m{iu}
              *{iu}{pb}u
```

- E.g.: \*{iu}{pb}{iu}, \*{iu}m{iu}

# Appendix: Results

- Word-final pattern:
  - 28 grammars: only \*m#
  - 1 grammar: only \*[nasal,labial]#
  - 3 grammars:

(44) the three runs (not counting run 23) at which the word-final restriction was not solely represented with the constraint \*m#

Run 12		Run 16			Run 17			
Constraint	Traditional notation	Weight	Constr aint	Traditional notation	Weight	Constraint	Traditional notation	Weight
*m#	*m#	2.29	*m#	*m#	2.27	*m#	*m#	2.15
*{aiu}m#	*Vm#	0.05 15	*{m}#	*[nasal,labial]#	0.16	*{aiu}m#	*Vm#	0.25