Learning allomorphy and segmentation*

Diana Apoussidou

University of Massachusetts Amherst

University of Amsterdam

d.apoussidou@uva.nl

A learnability problem

- How do you learn lexical/underlying representations (i.e. *hidden structure*) from surface forms?
- In e.g. phonology:
  - How do children acquire allomorphy?
  - How do children acquire the right segmentation?

Proposal

- Bootstrapping into the lexicon with morphemic and phonological information
- Lexical representations are encoded as constraints → Direct interaction with the grammar
- Lexical constraints are ranked as other constraints → No specific learning mechanism for acquiring the lexicon
- Acquisition of lexicon and grammar proceeds in an online fashion

Allomorphy

- English: *a pen* but *an oak*

→ allomorphs are listed in the lexicon

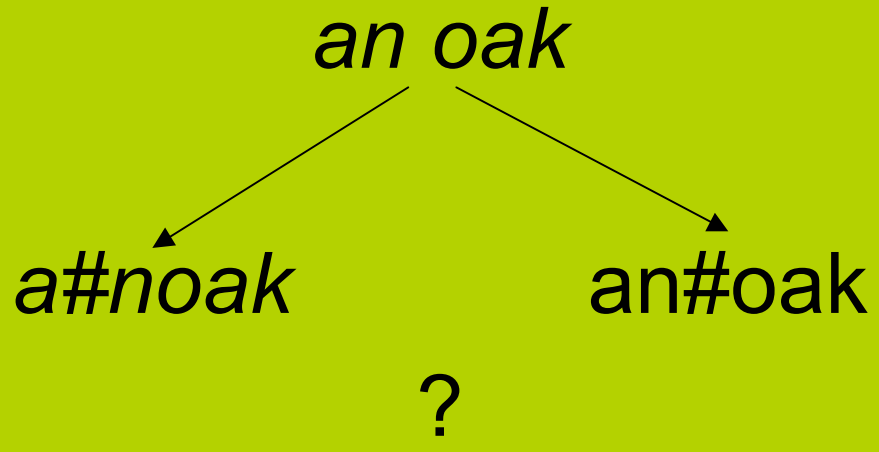
→ phonology chooses variant: *a* before consonants, *an* before vowels

Phonology gets to choose

	Onset	*C.C
☞ |a+pen| /a.pen/		
|an+pen| /an.pen/		*!

	Onset	*C.C
|a+oak| /a.ok/	*!	
☞ |an+oak| /a.nok/		

Segmentation



Help from paradigmatic information

a pen → |a+pen|

an oak → |a+nok|

big pen → |big+pen|

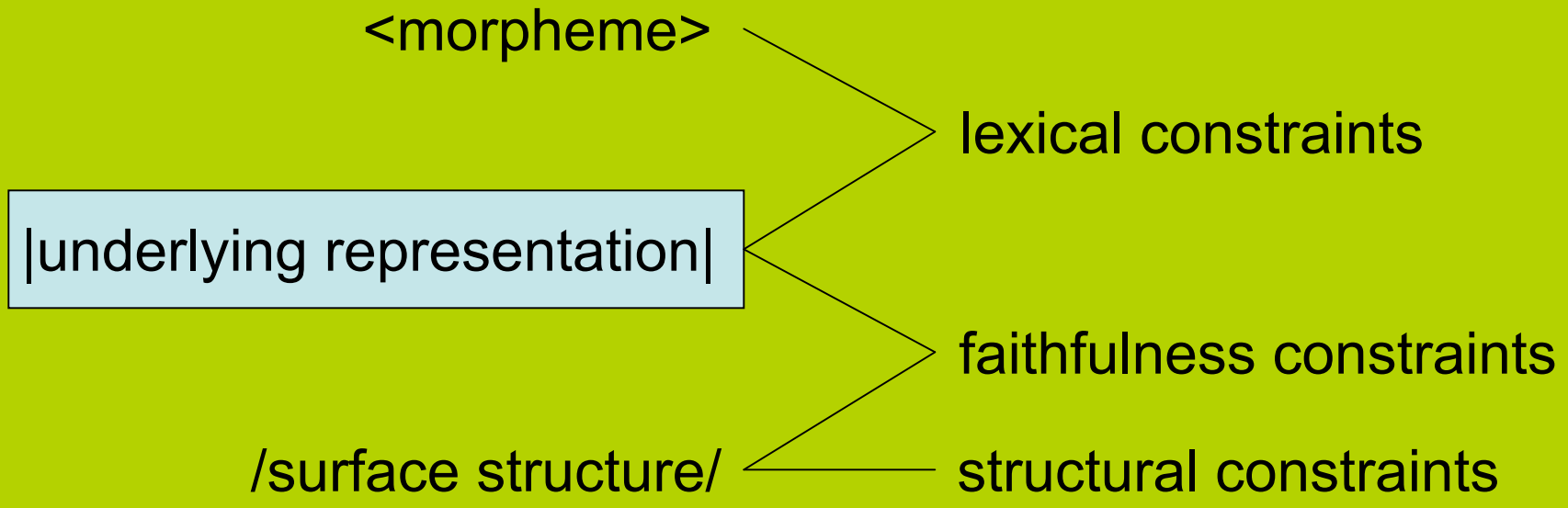
big oak → |big+ok|

Problem with Lexicon Optimization

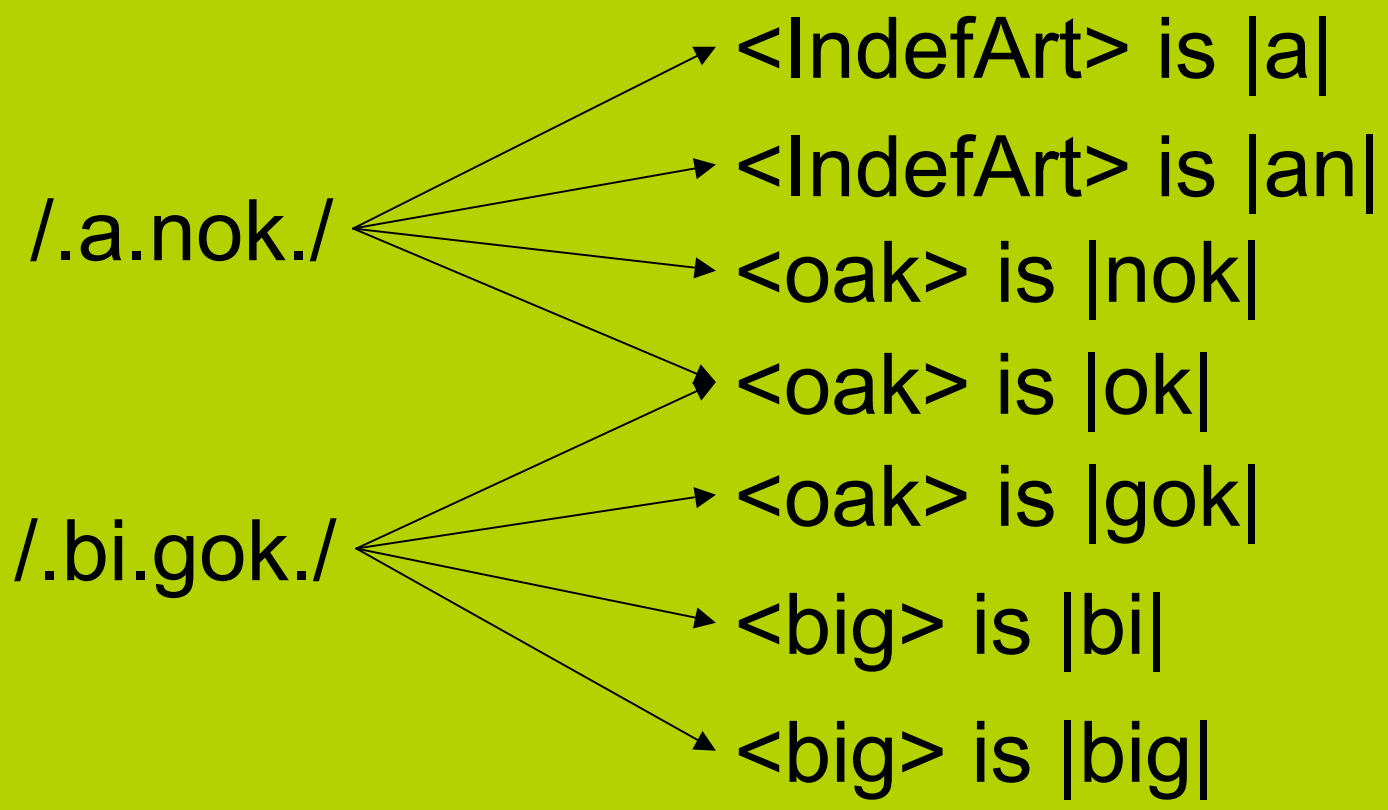
(Tesar & Smolensky 2000)

		Faith	Onset	*C.C
|an+ok|	/.a.nok./	√	√	√
|a+nok|	/.a.nok./	√	√	√
|big+ok|	/.bi.gok./	√	√	√
|bi+gok|	/.bi.gok./	√	√	√

The model


<morpheme>
|underlying representation|
/surface structure/
lexical constraints
faithfulness constraints
structural constraints

Positive lexical constraints



A possible learning path

/.a.nok./: |an+oak| or |a+noak|?

Comprehension

<IndefArt+oak> /.a.nok./	<IndefArt> |a|	<IndefArt> |an|	<oak> |ok|	<oak> |nok|	*C.C	Faith	Onset
👂 |a+noak| /.a.nok./		*	*				
|an+oak| /.a.nok./	*!			*			

Virtual production:

<IndefArt+oak>	<IndefArt> |a|	<IndefArt> |an|	<oak> |ok|	<oak> |nok|	*C.C	Faith	Onset
👄 |a+oak| /.a.ok./		*		←*			←*
👂 |a+noak| /.a.nok./		*	*!→				
|an+oak| /.a.nok./	*!			*			

A possible learning path

/.bi.gok./: |big+oak| or |bi+goak|?

<big+oak> /.bi.gok./	<big> |big|	<big> |bi|	<oak> |nok|	<oak> |ok|	*C.C	Onset	Faith
|bi+goak| /.bi.gok./	*!		*	*!			
👂 |big+oak| /.bi.gok./		*	*				

Producing <big+oak>:

<big+oak>	<big> |big|	<big> |bi|	<oak> |nok|	<oak> |ok|	*C.C	Onset	Faith
|bi+goak| /.bi.gok./	*!		*!	*			
👂 |big+oak| /.bi.gok./		*	*!→				
👄 |big+noak| /.big.nok./		*		←*	←*		

Ingredients to modelling

- Constraint set (faithfulness, structural and lexical)
- Training data (pairs of surface forms and morphemic information)
- Learning algorithm (gradual)

Constraints

- 1 Faith
- 2 structural constraints: Onset, *C.C
- 9 lexical constraints:

<oak> |ok|	<pen> |pen|
<oak> |nok|	<pen> |gpen|
<oak> |gok|	
<IndefArt> |a|	<big> |big|
<IndefArt> |an|	<big> |bi|

Training data and target forms

<IndefArt+pen> /.a.pen./ → |a+pen|

<IndefArt+oak> /.a.nok./ → |an+ok|

<big+pen> /.big.pen./ → |big+pen|

<big+oak> /.bi.gok./ → |big+ok|

Settings

- Praat (Boersma & Weenink 2007/2008)
- GLA, Stochastic OT
- OTMulti default settings
- Initial randomized ranking
- 400 000 pieces of training data, randomized
- Decreasing plasticity, evaluation noise
- 11 virtual learners

Results: 9 successful learners

IndefArt+oak |an+oak| /.a.nok./ >98.6%

IndefArt+oak |a+noak| /.a.nok./ <1.4%

IndefArt+pen |a+pen| /.a.pen./ >99.9%

big+oak |big+oak| /.bi.gok./ >99.9%

big+pen |big+pen| /.big.pen./ >99.9%

Final constraint ranking

	ranking value	*disharmony*	*plasticity*
<big> |big|	117.505	121.837	1.000000
FAITH	118.706	120.570	1.000000
ONSET	115.087	115.633	1.000000
<oak> |ok|	110.818	111.082	1.000000
*C.C	106.268	108.195	1.000000
<pen> |pen|	106.087	105.979	1.000000
<INDEFART> |a|	104.982	103.426	1.000000
<oak> |nok|	99.399	100.183	1.000000
<INDEFART> |an|	95.018	95.273	1.000000
<pen> |gpen|	93.913	93.334	1.000000
<oak> |gok|	89.783	92.760	1.000000
<big> |bi|	82.495	85.188	1.000000

an oak

	<INDEFART+oak>	FAITH	ONSET	<oak> |ok|	*C.C	<INDEFART> |a|	<oak> |nok|	<INDEFART> |an|	<oak> |gok|
		117.1	115.9	110.5	105.6	103.9	99.7	96.1	89.8
	<INDEFART+oak> |a+oak| /.a.ok./		*!				*	*	*
	<INDEFART+oak> |a+oak| /.a.ʔok./	*!					*	*	*
<1.4%	<INDEFART+oak> |a+noak| /.a.nok./			*!				*	*
	<INDEFART+oak> |a+noak| /.a.ok./	*!	*	*				*	*
	<INDEFART+oak> |a+goak| /.a.gok./			*!			*	*	
	<INDEFART+oak> |a+goak| /.a.ok./	*!	*	*			*	*	
>98.6%	☞ <INDEFART+oak> |an+oak| /.a.nok./					*	*		*
	<INDEFART+oak> |an+oak| /.a.ok./	*!	*			*	*		*
	<INDEFART+oak> |an+noak| /.an.nok./			*!	*	*			*
	<INDEFART+oak> |an+noak| /.a.nok./	*!		*		*			*
	<INDEFART+oak> |an+noak| /.a.ok./	*!*	*	*		*			*
	<INDEFART+oak> |an+goak| /.an.gok./			*!	*	*	*		
	<INDEFART+oak> |an+goak| /.a.gok./	*!		*		*	*		
	<INDEFART+oak> |an+goak| /.a.ok./	*!*	*	*		*	*		
	<INDEFART+oak> |bip+oak| /.bi.pok./					*	*	*!	*

Generalizations to *rat* and *apple*

IndefArt+*rat* |a+rat| /.a.ræt./ >99.9%

IndefArt+*apple* |an+æpl| /.a.næpl./ >99.9%

Results: 2 failed learners

<IndefArt+oak>	|an+oak| /.a.nok./
	|an+noak| /.an.nok./
	|an+noak| /.a.nok./
<IndefArt+pen>	|an+pen| /.a.pen./
	|an+pen| /.an.pen./

Results: 2 failed learners

<big+oak>	|big+oak| /.bi.gok./
	|big+noak| /.bi.nok./
	|big+noak| /.bi.gok./
	|big+noak| /.big.nok./
<big+pen>	|big+pen| /.big.pen./
	|big+pen| /.bi.pen./

Results: 2 failed learners

<IndefArt+apple> |an+apple| /.a.n\aepl./

<IndefArt+rat> |an+rat| /.an.r\aet./

|an+rat| /.a.r\aet./

A failed grammar

	ranking value	*disharmony*	*plasticity*
<big> |big|	106.816	106.816	1.000000
<oak> |nok|	105.880	105.880	1.000000
<pen> |pen|	105.492	105.492	1.000000
<INDEFART> |an|	104.893	104.893	1.000000
<oak> |ok|	100.622	100.622	1.000000
ONSET	100.000	100.000	1.000000
<INDEFART> |a|	95.107	95.107	1.000000
<pen> |gpen|	94.508	94.508	1.000000
<oak> |gok|	93.498	93.498	1.000000
<big> |bi|	93.184	93.184	1.000000
*C.C	91.800	91.800	1.000000
FAITH	89.874	89.874	1.000000

big pen

	<big+pen>	106.8 <big> |big|	105.5 <pen> |pen|	100.0 ONSET	94.5 <pen> |gpen|	93.2 <big> |bi|	91.8 *C.C	89.9 FAITH
24.7%	<big+pen> |big+pen| /.big.pen./				*	*	*!	
75.3% ☞	<big+pen> |big+pen| /.bi.pen./				*	*		*
	<big+pen> |big+gpen| /.big.gpen./		*!			*	*	
	<big+pen> |big+gpen| /.big.pen./		*!			*	*	*
	<big+pen> |bi+pen| /.bi.pen./	*!			*			
	<big+pen> |bi+gpen| /.big.pen./	*!	*				*	

Implications

- Lexicon as computation/constraint ranking → enables direct interaction with the grammar
- Acquiring a lexicon proceeds as acquiring the grammar → no specific learning mechanism needed (cf. Tesar et al. 2003)
- Proceeds online; lexicon and grammar can be acquired simultaneously
- One grammar and lexicon at a time (cf. Jarosz 2006)
- Positive lexical constraints reduce the number of possible underlying forms in learning

Problems

- Some learners fail
- Paradigmatic information does not only help, but can confuse learners

Selected references

Apoussidou, D. (2007). *The learnability of metrical phonology*. Dissertation, LOT Dissertation Series 148.

Boersma, P. & D. Weenink (2007/2008). Praat: doing phonetics by computer (Version 4.4-5.0.34). [Computer program]. www.praat.org.

Jarosz, G. (2006). Rich lexicons and restrictive grammars - Maximum likelihood learning in Optimality Theory. Dissertation, Baltimore, Maryland.

Prince, Alan & Paul Smolensky (1993). Optimality Theory: constraint interaction in generative grammar. Ms., Rutgers University, New Brunswick, NJ, and University of Colorado, Boulder. Technical Report RuCCS-TR-2. Rutgers Center for Cognitive Science.

Tesar, B. & P. Smolensky (2000). *Learnability in Optimality Theory*. MIT Press.

Tesar, B., J. Alderete, G. Horwood, N. Merchant, K. Nishitani & A. Prince (2003). Surgery in language learning. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (WCCFL), eds. G. Garding and M. Tsujimura, 477-490. Somerville, MA: Cascadilla Press. ROA 619.

Wolf, M. (2007). Lexical insertion occurs in the phonological component. To appear in Bernard Tranel (ed.), *Understanding allomorphy: Perspectives from Optimality Theory*. London: Equinox.

*Acknowledgements

- Thanks to Paul Boersma, Karen Jesney and Joe Pater for their comments
- Thanks to the Netherlands Organisation for Scientific Research (NWO) for funding this (Rubicon grant 446-07-030)