Rule-based learning of phonological optionality and opacity

Ezer Rasin, Iddo Berger, Nur Lan and Roni Katzir

MIT and Tel Aviv University

NELS 48, October 27–29, 2017, University of Iceland

The problem

Optionality and opacity pose obvious challenges for the child learning the phonology of their ambient language

Optionality in French (Dell 1981)

(1) a. tabl ∼ tab 'table'

b. parl ∼ *par 'speak'

(2) $L \rightarrow \emptyset$ /[−*son*]__# (optional)

Counterfeeding opacity in Catalan (Mascaró 1976)

(3) a. kuzí ∼ kuzín-s 'cousin.SG ∼ cousin.PL'

b. kələ́n ∼ kələ́nt-ə 'hot.MASC ∼ hot.FEM'

(4) Ordered rules (simplified):

1. $N \rightarrow \emptyset$ /__#
2. $C \rightarrow \emptyset$ /N__#

State of the art and our contribution

- Children acquire optionality and opacity in a variety of languages (Dell 1981, McCarthy 2007)
- No learners in the literature can handle optionality or opacity distributionally, from unanalyzed input data alone

We show how optional processes and opaque interactions (including both counterfeeding and counterbleeding) can be acquired distributionally using the principle of Minimum Description Length (MDL; Solomonoff 1964, Rissanen 1978)

The MDL criterion

Evaluation criterion that balances two competing factors:

- The simplicity of the grammar, $|G|$
- The tightness of fit of the grammar to the data, $|D : G|$

MDL evaluation metric

$$\underbrace{\underbrace{0101011010101001010}_{\textit{Lexicon}}\underbrace{10101010010}_{\textit{Rules}}}_{G}\underbrace{10100010110101}_{D:G}$$

- We use an adaptation of Rasin & Katzir's (2016) MDL learner (originally used for OT phonology) to rule-based phonology

Properties of our learner

Distributional learning

The learner is exposed to unanalyzed input data alone

- No paradigms or morphological information
- No URs
- No information about the phonological mapping

General learning criterion

Acquires a segmented lexicon and the phonology in a simple, unified way

Interactions

Can handle interactions of different phenomena (e.g., nasal deletion and cluster simplification in Catalan)

Criterion I: Grammar economy; leads to overgeneration

SPE evaluation metric (Chomsky and Halle 1968)

If G and G' can generate the data D, and if $|G| < |G'|$, prefer G to G'

Input data (D)

tabl, tab, arbr, arb, parl, ... (no par)

Hypothesis G_1 (complex)

- Lex: /tabl/, /tab/, /arbr/, /arb/, /parl/, ...
- Rules: $\emptyset$

Hypothesis G_2 (simple, overgenerating)

- Lex: /tabl/, /arbr/, /parl/, ...
- Rules: $L \rightarrow \emptyset$ (optional) (cf. $L \rightarrow \emptyset$ /[$-son$]__#)

Overgeneration: the child will rule in *[par] for /parl/

Criterion II: Subset Principle; leads to overfitting

Subset evaluation metric (Dell 1981; Berwick 1985)

If G and G' can both generate the data D, and if the language of G is a proper subset of the language of G', prefer G to G'

Input data (D)

tabl, tab, arbr, arb, parl, sabl, ... (no sab)

Hypothesis G_3 (correct hypothesis)

- Lex: /tabl/, /arbr/, /parl/, /sabl/, ...
- Rules: $L \rightarrow \emptyset$ /[$-son$]__# (optional) Generates [sab] for /sabl/

Hypothesis G_1 (complex, overfitting)

- Lex: /tabl/, /tab/, /arbr/, /arb/, /parl/, /sabl/, ...
- Rules: $\emptyset$

Undergeneration: the child will rule out [sab] for /sabl/

The MDL principle

MDL evaluation metric

If G and G' can both generate the data D, and if $|G| + |D : G| < |G'| + |D : G'|$, prefer G to G'

$$\underbrace{\underbrace{0101011010101001010}_{\textit{Lexicon}}\underbrace{10101010010}_{\textit{Rules}}}_{G}\underbrace{10100010110101}_{D:G}$$

- Balances grammar economy and restrictiveness
- Restrictiveness is cashed out in terms of simplicity
- Rissanen (1978), building on the work of Solomonoff (1964), Kolmogorov (1965), and Chaitin (1966)
- Helpful across a range of grammar induction tasks in works such as Horning (1969), Berwick (1982), Stolcke (1994), Grünwald (1996), de Marcken (1996), Clark (2001), Goldsmith (2001), and Goldwater and Johnson (2004)

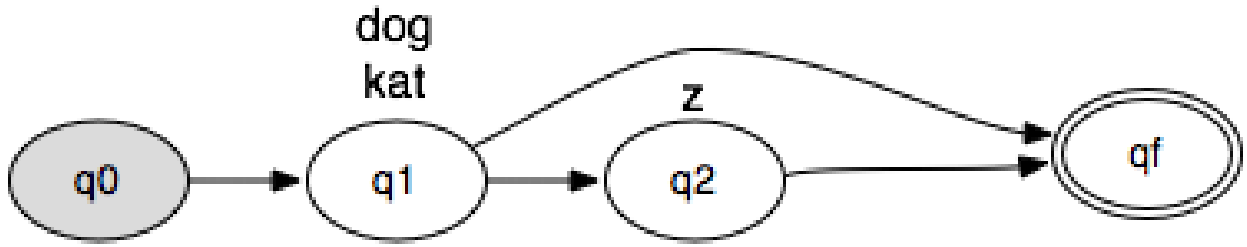
Representations

Segments

Represented as bundles of binary features

Lexicon (Hidden Markov Model)

- Contains URs of morphemes
- Contains information about how URs can be combined



Phonological rules

A list of ordered rules of the form:

$$\underbrace{A}_{\text{focus}} \rightarrow \underbrace{B}_{\text{change}} / \underbrace{X}_{\text{left context}} \_ \underbrace{Y}_{\text{right context}} \quad \text{(optional?)}$$

Measuring $|G|$ (illustration)

Vowel harmony: textbook notation

$$[-cons] \rightarrow [-back] \,/\_\_\, [+cons]^* \begin{bmatrix} -cons \\ -back \end{bmatrix} \text{ (optional)}$$

Vowel harmony: string notation

$$-cons\#_{rc} - back\#_{rc}\#_{rc} + cons * \#_b - cons\#_f - back\#_{rc}1\#_{rc}$$

Symbol	Code	Symbol	Code
$\#_f$ (feature)	0000	cons	0110
$\#_b$ (bundle)	0001	voice	0111
$\#_{rc}$ (rule component)	0010	velar	1000
+	0011	back	1001
-	0100	...	...
*	0101	...	...

More information in Rasin, Berger, Lan and Katzir (2017), available at http://ling.auf.net/lingbuzz/003665

Measuring $|D : G|$ (illustration)

Encoding length of [parl] given the simple rule: 4 bits

- Lex: /tabl/, /arbr/, /parl/
- Rules: $L \rightarrow \emptyset$ (optional)

Step I: Choose UR: /parl/ (2 bits)
Step II: Apply optional rule? no (2 bits – 1 bit for each of /r/, /l/)

Encoding length of [parl] given the complex rule: 2 bits

- Lex: /tabl/, /arbr/, /parl/
- Rules: $L \rightarrow \emptyset$ /[$-son$]__# (optional)

Step I: Choose UR: /parl/ (2 bits)
Step II: -

Summary

G_1 (chosen by grammar economy alone; overly general)

$$\underbrace{\underbrace{010110010001}_{\textit{Lex}=/tabl/,/arbr/,/parl/,\ldots}\ \underbrace{1011101001}_{\textit{Rules}=L\to\emptyset}}_{G}\ \underbrace{\underbrace{010}_{4}\ \underbrace{0}\ \underbrace{110}_{3}\ \underbrace{011}_{4}\ \underbrace{0}\ \underbrace{011}_{4}\ \underbrace{0}\ \underbrace{010}_{4}\ \underbrace{0}\ldots}_{D:G}$$

G_2 (chosen by restrictiveness alone; overfitting)

$$\underbrace{\underbrace{11010011010011101100001011101}_{\textit{Lex}=/tabl/,/tab/,/arbr/,/arb/,/parl/\ldots}\ \underbrace{10}_{\textit{Rules}=(none)}}_{G}\ \underbrace{\underbrace{010}_{3}\ \underbrace{110}_{3}\ \underbrace{011}_{3}\ \underbrace{011}_{3}\ \underbrace{010}_{3}\ldots}_{D:G}$$

G_3 (chosen by MDL; balanced)

$$\underbrace{\underbrace{010110010001}_{\textit{Lex}=/tabl/,/arbr/,/parl/,\ldots}\ \underbrace{1001101110110}_{\textit{Rules}=L\to\emptyset}\ /[-son]\_\_\#}_{G}\ \underbrace{\underbrace{010}_{4}\ \underbrace{0}\ \underbrace{110}_{3}\ \underbrace{011}_{4}\ \underbrace{1}\ \underbrace{011}_{3}\ \underbrace{010}_{3}\ldots}_{D:G}$$

1. Introduction

2. The MDL criterion

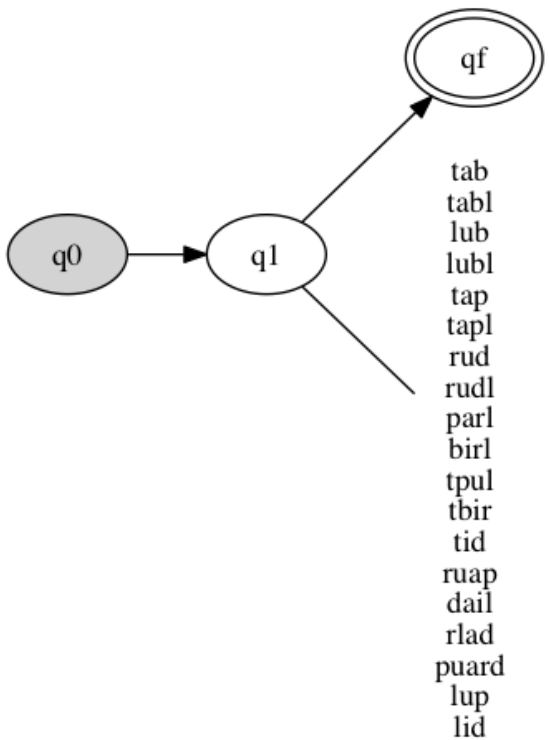
3. Simulations
   - French optionality
   - Counterfeeding opacity in Catalan

Simulation I: French optionality

Data

tab, tabl, parl, lub, lubl, tap, tapl, rud, rudl, birl, dail, lid, lup, puard, rlad, ruap, tbir, tid, tpul

$$G_{initial} = \begin{cases} \textsc{Rules}: \quad \emptyset \\ \textsc{Lex}: \end{cases}$$



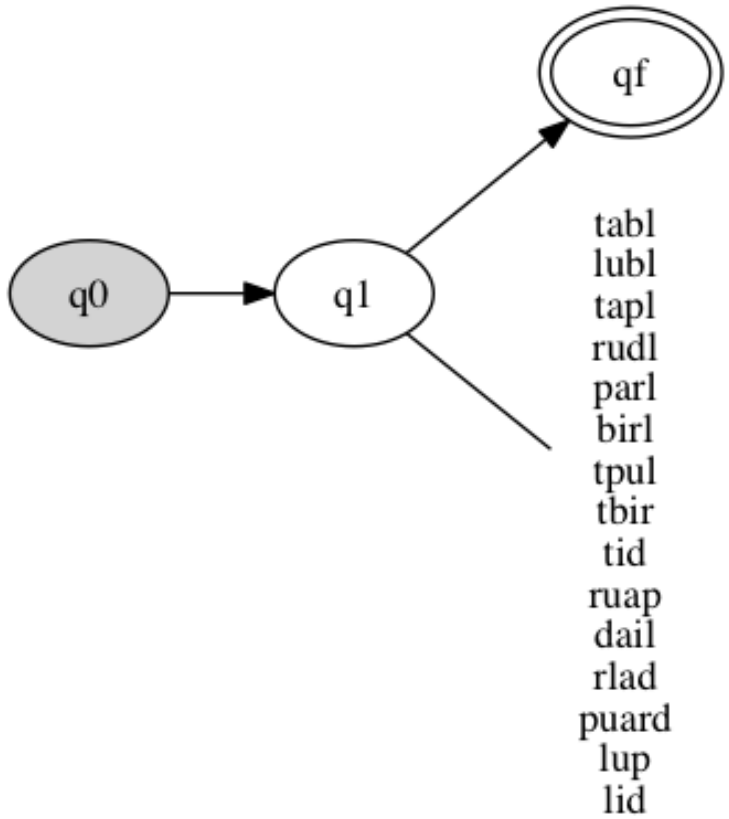
Description length: $|G_{initial}| + |D{:}G_{initial}| = 7,740 + 4,750 = 12,490$

Simulation I: French optionality: result

$G_{final} =$ {

RULES: $R_1 : [+liquid] \rightarrow \emptyset / [-son]\_\_(optional)$

LEX:



}

Description length: $|G_{final}| + |D{:}G_{final}| = 6,493 + 4,200 = 10,693$

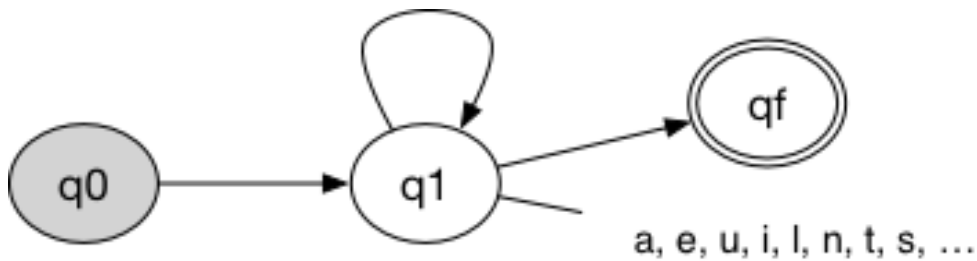
Simulation II: Counterfeeding opacity in Catalan

Data

55 words created by taking all combinations of 11 stems with 5 suffixes and applying nasal deletion and cluster simplification, in this order

stem\suffix	∅	-s	-et	⋯
kalent	kalen	kalents	kalentet	
kuzin	kuzi	kuzins	kuzinet	
⋯				

$$G_{initial} = \begin{cases} \textsc{Rules}: \quad \emptyset \\ \textsc{Lex}: \end{cases}$$



Description length: $|G_{initial}| + |D{:}G_{initial}| = 145 + 44,625 = 44,770$

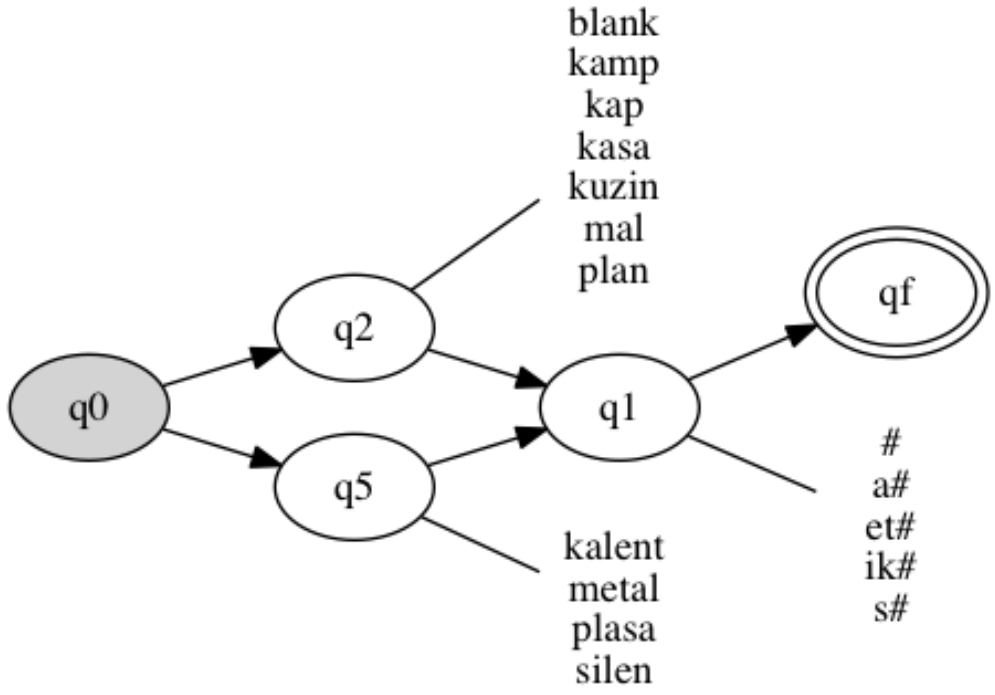
Simulation II: Counterfeeding opacity in catalan: result

$G_{final} =$

RULES: $R_1 : [+nasal] \rightarrow \emptyset /$ ____ $\#$

$R_2 : [+cons] \rightarrow \emptyset / [+nasal]$ ____ $\#$

LEX:



Length : $|G_{final}| + |D{:}G_{final}| = 451 + 8,825 = 9,276$

Other results: comparison of bleeding and counterbleeding

Two simulations: same lexicon and rules, but opposite rule ordering

Bleeding: Vowel epenthesis and voicing assimilation in English
Counterbleeding: Reversed English

- Learning succeeds with both orderings
- No difference in terms of the MDL score
- Simulation with opacity took much longer to converge

Implications

- Optionality and opacity can be learned from distributional evidence alone
- Further support for the MDL metric which is general and is not designed with optionality or opacity (or even phonology) in mind
- The same metric that supports e.g., segmentation allows us to handle the challenge of acquiring optionality and opaque phonological interactions

Acknowledgements

Thanks to Adam Albright, Naomi Feldman, Michael Kenstowicz, Donca Steriade, and the audience at MIT.

Bibliography I

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral Dissertation, MIT, Cambridge, MA.

Berwick, Robert C. 1985. *The acquisition of syntactic knowledge*. Cambridge, Massachusetts: MIT Press.

Chaitin, Gregory J. 1966. On the length of programs for computing finite binary sequences. *Journal of the ACM* 13:547–569.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row Publishers.

Clark, Alexander. 2001. Unsupervised language acquisition: Theory and practice. Doctoral Dissertation, University of Sussex.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.

Bibliography II

Goldwater, Sharon, and Mark Johnson. 2004. Priors in Bayesian learning of phonological rules. In *7th Annual Meeting of the ACL Special Interest Group on Computational Phonology*, 35–42.

Grünwald, Peter. 1996. A minimum description length approach to grammar inference. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. G. S. S. Wermter and E. Riloff, Springer Lecture Notes in Artificial Intelligence, 203–216. Springer.

Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.

Kolmogorov, Andrei Nikolaevic. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission (Problemy Peredachi Informatsii)* 1:1–7. Republished as Kolmogorov (1968).

Bibliography III

Kolmogorov, Andrei Nikolaevic. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2:157–168.

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT, Cambridge, MA.

Rasin, Ezer, Iddo Berger, and Roni Katzir. 2017. Learning rule-based morpho-phonology. Ms., MIT and TAU.

Rasin, Ezer, and Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry* 47:235–282.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.