

Restrictive Learning with Distributions over Underlying Representations

Karen Jesney, Joe Pater & Robert Staubs

University of Massachusetts Amherst

{kjesney, pater, rstaubs} @ linguist.umass.edu

Workshop on Computational Modeling of Sound Pattern
Acquisition – Edmonton, AB – February 13, 2010

Acknowledgements

This research has been done partially in collaboration with Diana Apoussidou and David Smith. It was supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts, Amherst. Thank you to the audiences at NECPhon 2009, and the LSA for feedback.

1. Overview

In acquiring a language, a child must establish:

➤ **a grammar** that generates the set of forms permitted by the target language

e.g., If the language has regular final devoicing, the grammar should limit the forms produced to those that follow this restriction.

➤ **a set of underlying representations** that associate meanings to phonological underlying forms

This presents a challenge – especially in cases when the surface form of a morpheme varies based on the phonological context.

- choosing a UR affects the choice of grammar, and vice versa

Our approach:

- Kager (2009) proposes OT “allomorphy” as an account of phenomena dealt with in terms of abstract URs and lexically specific constraints. We extend this account to one of learning a distribution over URs – cf. exemplar models, e.g., Pierrehumbert 2001, 2002, 2003.
- We implement this using constraints on URs as proposed by Boersma (1999) and Apoussidou (2007), reformalized along lines similar to Eisenstat (2009)

Our big picture points:

1. A distinction may not need to be drawn between learning allomorphy and “regular” URs, so long as biases are in place to ensure restrictiveness.

➤ minimizing the weights of Faithfulness constraints

2. A learner may not need to search a space of non-surface-existing URs.

➤ constraints associating meanings with observed surface forms are adequate for some cases claimed to require abstract URs.

3. With a model of grammar that yields variation, distributions over URs extend to cases that abstract URs cannot deal with.

Outline:

§2 Interactions between URs and the grammar:

- constraints on URs
- how distributions over URs yield “abstract URs”

§3 Learning model and simulation results

§4 Some consequences:

- exceptionality
- lexically-conditioned variation

§5 Conclusions

2. Interactions between URs and the grammar

Many languages show the following type of alternation in the phonological form of morphemes:

- (1) [mat] “bush”
[mada] “bushes”

Assuming the plural morpheme is /-a/, “bush” has two forms:

- (2) [mat] [mad]

A much-discussed hidden structure problem: which is the underlying (lexical) form?

- (3) /mat/ *or* /mad/?

Our learner's answer is **both**: the problem becomes one of finding appropriate weights on constraints favoring each one.

Constraints on URs – i.e., Lexical constraints (Boersma 1999, Apoussidou 2007, Eisenstat 2009):

(4) BUSH → /mat/

Assign a score of -1 if the UR is not /mat/

BUSH → /mad/

Assign a score of -1 if the UR is not /mad/

Hypothesis:

- These constraints are *positively-formulated*,
- They are limited to the set of surface allomorphs observed in the target language data.

(5)

“bush”	*CODA-VOICE	IDENT-VOICE	BUSH →/mad/	BUSH →/mat/
/mat/→[mat]			-1	
/mad/→[mat]		-1		-1
/mad/→[mad]	-1			-1


The grammar defines a **probability distribution** over candidates.

- Selection of an optimum depends upon this distribution
- The learning goal for the tableau in (5):

(6) $p(/mat/→[mat]) + p(/mad/→[mat]) = 1$

An example (for now, highest score gets $p = 1$):

(7)

	*CODA- VOICE $W = 1$	IDENT- VOICE $W = 1$	BUSH →/mad/ $W = 1$	BUSH →/mat/ $W = 10$	H
“bush”					
 /mat/→ [mat]			-1		-1
/mad/→ [mat]		-1		-1	-11
/mad/→ [mad]	-1			-1	-11

- The high weight of BUSH→/mat/ causes the UR /mat/ to be strongly preferred.
- The candidate /mat/→[mat] is selected because it violates no markedness or faithfulness constraints.

This solution will not necessarily be appropriate given other data in the language.

➤ e.g., if it is a “final devoicing” language:

(8) [mat]	“bush”	[bat]	“tree”
[mad+a]	“bushes”	[pata]	“flower”


The problem (given the standard constraint set):

➤ By making the weight of BUSH→/mat/ so much higher than the weight of BUSH→/mad/, the learner is forced to treat [mad+a] as an instance of intervocalic voicing. However, this solution which is inconsistent with [pata].

A non-standard solution:


- Allow freer selection between the two underlying forms

(9)

	*CODA-VOICE $W = 4$	IDENT-VOICE $W = 3$	BUSH →/mad/ $W = 1$	BUSH →/mat/ $W = 1$	<i>H</i>
"bush"					
 /mat/→ [mat]			-1		-1
/mad/→ [mat]		-1		-1	-4
/mad/→ [mad]	-1			-1	-5

- UR /mat/ selected in the bare form to minimize violations of *CODA-VOICE and IDENT-VOICE.

(10)

	IDENT-VOICE $w = 3$	*VTV $w = 2$	BUSH →/mad/ $w = 1$	BUSH →/mat/ $w = 1$	H
“bushes”					
/mat+a/→ [mata]		-1	-1		-3
/mad+a/→ [mata]	-1	-1		-1	-6
/mat+a/→ [mada]	-1		-1		-4
 /mad+a/→ [mada]				-1	-1

➤ UR /mad/ selected in the derived form to minimize violations of *VTV and IDENT-VOICE.

This is a non-standard analysis because the UR selected depends upon phonological context.

- This type of analysis is typically reserved for cases of suppletion with a partially phonologically predictable element (e.g., a/an).


We take this non-standard solution to be viable because with $w(*\text{CODA-VOICE}) > w(\text{IDENT-VOICE})$ it passes the “Richness of the Base” test (Prince & Smolensky 1993/2004).

- In a language where final devoicing applies across the board, this pattern should be independent of the UR selected – i.e., our learner should learn that final consonants are predictably voiceless.

The test:

- Assume the learner is exposed to the new word [bada] “flowers”, seen only in its derived form.
- It segments [-a] as an affix, and assigns [bad] with the meaning “flower”.
- The only observed form is [bad] so the learner infers /bad/ as the UR for “flower”.
- *Does [bad] or [bat] surface in underived contexts?*

(11)

“flower”	*CODA- VOICE $W = 4$	IDENT- VOICE $W = 3$	H
 /bad/ → [bat]		-1	-3
/bad/ → [bad]	-1		-4

This approach is also successful in more complex cases.

➤ Tesar's (2006) Paka language was designed a test of theories of UR learning.

(12)

	/re-/	/ri:-/	/'ro-/	/'ru:-/
/-se/	'rese	'ri:se	'rose	'ru:se
/-'si/	re'si	ri'si	'rosi	'ru:si
/-'so:/	re'so:	ri'so:	'roso	'ru:so


Tesar's "paka" language: stressed syllables preceded by single quote (e.g. 'ro), long vowels indicated by colon (e.g. ri:).

The UR /ri:-/ does not exist in any surface form – it surfaces as *short unstressed* [ri] or *long stressed* ['ri:]. In a theory with single URs it is required because:

- (13)
- It is underlyingly *long* in contrast with /re-/
 - It is underlyingly *stressless* in contrast with /'ru:-/


In the allomorphic analysis, the morpheme has two URs, /'ri:/ and /ri/.

(14)

RI+SE	STRESS- FAITH $w = 3$	MAIN- LEFT $w = 2$	RI→ /'ri:/ $w = 1$	RI→ /ri/ $w = 1$	<i>H</i>
 /'ri:+se/→ ['ri:se]				-1	-1
/ri+se/→ ['rise]	-1		-1		-4
/ri+se/→ [ri'se]	-1	-1	-1		-6

➤ When it combines with stressless /se/, stress faithfulness prefers the UR /'ri:/.

(15)


RI+'so:	STRESS- FAITH $w = 3$	MAIN- LEFT $w = 2$	RI→ /'ri:/ $w = 1$	RI→ /ri/ $w = 1$	<i>H</i>
/'ri:+'so:/→ ['ri:so]	-1			-1	-4
/ri+'so:/→ ['riso]	-2		-1		-7
 /ri+'so:/→ [ri'so:]		-1	-1		-3

➤ When it combines with long stressed /-'so:/, stress faithfulness instead prefers the UR /ri/.

In the case of morphemes that do not alternate, violation of STRESS-FAITH is unavoidable.

➤ e.g., the morpheme with 'RU: has a single UR

(16)

	STRESS- FAITH $w = 3$	MAIN- LEFT $w = 2$	'RU:→ /ru:/ $w = 1$	<i>H</i>
 /'ru:+'so:/→ [ru:so]	-1			-3
/'ru:+'so:/→ [ru'so:]	-1	-1		-5

The next issue: establishing appropriate weights for the constraints in order to ensure restrictiveness

3. Learning model and simulation results

Rather than always selecting the candidate form with the highest Harmony, in simulations we compute a probability distribution over candidates.

- The probability of an output candidate y_{ij} given an input x_i is the exponential of its harmony normalized by the sum of the exponentials of the corresponding input:

$$(17) \quad p(y_{ij} \mid x_i) = \frac{1}{Z_i} e^{H_{ij}}$$
$$Z_i = \sum_{j'} e^{H_{ij'}}$$

- This is the definition of probability used in Maximum Entropy OT (Goldwater & Johnson 2003, Wilson 2006)

However, a given overt form y_{ij} may correspond to a number of different possible underlying representations z_{ijk} .

➤ e.g., candidates /bad/→[bat] and /bat/→[bat]

➤ The probability assigned to a given overt form y_{ij} is the sum of the probabilities of all full structures consistent with it:

$$(18) \quad p(y_{ij} \mid x_i) = \sum_k p(y_{ij}, z_{ijk} \mid x_i)$$
$$= \frac{1}{Z_i} \sum_k e^{H_{ijk}}$$
$$Z_i = \sum_{j'k'} e^{H_{ij'k'}}$$

We can then learn appropriate weights w^* for the constraints are learned by **maximizing the log likelihood** of the training data, as in (19).

$$(19) \quad w^* = \max_w \sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*}$$

➤ Without hidden structure this is Maximum Entropy learning.

➤ A similar approach to UR learning (with single URs) was developed by Eisenstat (2009). Jarosz (2006) develops a distinct approach to UR learning with maximum likelihood.

Unconstrained, weights will tend toward infinity to maximize the probability of the observed forms.

To enforce convergence we introduce an L_2 (Gaussian) **regularization** term:

$$(20) \quad w^* = \max_w \left[\sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} - \frac{1}{2\sigma^2} \sum_\ell w_\ell^2 \right]$$

Regularization can also prevent the learner from becoming trapped in local maxima.

We also establish a hard minimum of 0.0 on constraint weights to prevent “beneficial” violations.

Problem: The learner can do well on the objective function by merely memorizing the correct forms – i.e., by weighting Faithfulness highly.

- The resulting grammar will not be restrictive.
- We thus enforce a simple $M > F$ bias, following e.g., Hayes (2004), Prince & Tesar (2004), Smolensky (1996).
- To do this, we maximize the difference between the sums of the two classes of constraints: Markedness and Lexical constraints vs. Faithfulness constraints.
- This gives an approximation to Prince & Tesar's (2004) maximization of R-measure.

The factor λ controls the weight of the term.

Combined with regularization, this bias keeps the weights of Faithfulness constraints as low as possible and the weights of other constraints as high possible while maintaining consistency with the target data (cf. Jesney & Tessier to appear).

$$(21) \quad w^* = \max_w \sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} - \frac{1}{2\sigma^2} \sum_\ell w_\ell^2 + \lambda \left[\sum_{w_m \in \mathcal{M}} w_m - \sum_{w_f \in \mathcal{F}} w_f \right]$$

We tested this approach using Tesar's (2006) Paka language described in the previous section.

MAINSTRESSLEFT	Stress is on the leftmost syllable
MAINSTRESSRIGHT	Stress is on the rightmost syllable
WEIGHTTOSTRESS	Long vowels are stressed
*V:	Vowels are short
IDENTSTRESS	Corresponding input and output vowels have identical stress
IDENTLENGTH	Corresponding input and output vowels have identical length
/re/, /'re/, ...	Constraints on underlying representations

Simulation:


Initial weights set at 1.0, $\sigma^2 = 48.0$, $\lambda = 0.3$

17.19	/so:/	13.65	/re/
16.51	/ri/	12.29	/ri:/
15.15	/re/	11.61	/so/
14.88	MAINSTRESSLEFT	10.99	/si/
14.40	/ro/	7.41	/si/
14.40	/ru:/	6.26	IDENTSTRESS
14.40	/se/	2.56	IDENTLENGTH
14.40	WEIGHTTOSTRESS	0.00	*V:
13.92	MAINSTRESSRIGHT		

One case potentially requiring an abstract UR:

- RI+SE → ['ri:se], *['rise], *[ri'se]
- Here, selecting the UR /'ri:/ allows stress to be placed without violating IDENT-STRESS.


(22)

RI+SE	RI→ /ri/	MAIN- LEFT	MAIN- RIGHT	RI→ /'ri:/	IDENT- STRESS	<i>H</i>
	16.51	14.88	13.92	12.29	6.26	
 /'ri:+se/→ ['ri:se]	-1		-1			-30.43
/ri+se/→ ['rise]			-1	-1	-1	-32.47
/ri+se/→ [ri'se]		-1		-1	-1	-33.43

Another case potentially requiring an abstract UR:

- RI+'SO: → [ri'so:], *['riso], *['ri:so]
- Here, selecting the UR /ri/ allows stress to be placed without violating IDENT-STRESS.

(23)

RI+'SO:	RI→ /ri/	MAIN- LEFT	MAIN- RIGHT	RI→ /'ri:/	IDENT- STRESS	<i>H</i>
	16.51	14.88	13.92	12.29	6.26	
/'ri:+'so:/→ ['ri:so]	-1		-1		-1	-36.69
/ri+'so:/→ ['riso]			-1	-1	-2	-38.73
 /ri+'so:/→ [ri'so:]		-1		-1		-27.17

These results also pass a RotB test.

- When all possible combinations of URs are supplied, the surface forms generated follow the patterns of the target language.
- The target language has no unstressed long vowels; the weights learned should preserve this pattern under any combination of URs.
- The highest probability these weights give to an unstressed long vowel is **6.75×10^{-6}** .

4. Some consequences

A model with constraints on URs allows certain cases of exceptionality to be modeled in a straightforward fashion.

In this model, non-alternating forms (exceptional or not) have only a single UR available. Alternating forms have multiple URs (Kager 2009).

➤ e.g., a Turkish-like language that has regular final devoicing, and also includes words that maintain a final voiced consonant.

(24)	[mat]	“bush”	[mada]	“bushes”
	[pat]	“tree”	[pata]	“trees”
	[bad]	“flower”	[bada]	“flowers”

There are essentially three types of consonant in this system
– the contrast is sometimes captured using archiphonemes
(Inkelas, Orgun & Zoll 1997)


(25)	Alternating /T/	[0voice]
	Non-alternating voiceless /t/	[-voice]
	Non-alternating voiced /d/	[+voice]

Instead, in this model we can rely on the available URs and lexical constraints.

(26)	BUSH→/mat/,	BUSH→/mad/
	TREE→/pat/	
	FLOWER→/bad/	


If IDENT has sufficient weight and only a single UR is available, the voicing will emerge in all environments, including in word-final position.

(27)

FLOWER	IDENT-VOICE $w = 3$	*CODA-VOICE $w = 2$	FLOWER →/bad/ $w = 2$	<i>H</i>
/bad/→ [bat]	-1			-3
 /bad/→ [bad]		-1		-2

When multiple URs are available, however, markedness constraints can choose between them, thus yielding alternation.

(28)

	IDENT-VOICE $w = 3$	*CODA-VOICE $w = 2$	BUSH →/mad/ $w = 1$	BUSH →/mad/ $w = 1$	<i>H</i>
 /mat/→ [mat]			-1		-1
/mad/→ [mat]	-1			-1	-4
/mad/→ [mad]		-1		-1	-3

A model with that encodes probabilities over UR choices also allows us to handle cases of lexically-conditioned variation that escape abstract URs – *cf.* Pierrehumbert 2002.

- E.g., French “schwa” is often analyzed using abstract segment.
- Some schwas occasionally delete, some schwas never delete, and some clusters never include schwa.

(29)	la s(e)maine	‘the week’
	le m(e)lon	‘the melon’
	la belon	‘the oyster (a particular kind)’
	la blonde	‘the blonde’
	le SMIC	‘the unemployment insurance’

Analysis in terms of abstract URs:

- (30) Alternating vowel “V” – underspecified vowel
Fixed vowel /ə/ (/œ/,/ø/) – fully specified vowel
Absence of vowel ∅

However, there is not only a distinction between words that alternate and those that don't. Words that alternate differ in the probability of deletion – for discussion, see Coetzee & Pater to appear, Pater 2008.

- (31) le s(e)mestre low probability of deletion
 la s(e)maine high probability of deletion

There is plenty of evidence for this empirical claim:

- (32) a. Dictionaries find the two-way categorization of alternating/non-alternating inadequate (Walker 1996)
- b. Corpus-based studies note that some words show greater frequency of deletion than others (e.g., Hansen 1994, Eychenne 2007, Eychenne & Pustka 2007)
- c. Racine & Grosjean (2002) provide data from a production study showing a wide range of deletion frequencies across words
- d. Racine (2007) shows speakers can judge relative deletability

Weights on URs can be used to encode this sort of lexical conditioning.

This lexical conditioning interacts with a variety of phonological factors – e.g., sequences of three consonants are generally avoided (see Dell 1973 for discussion).

Our learner achieves considerable success in matching realistic probabilities of deletion for different words.

(33)

	target probability of deletion	learned probability of deletion
le belon	0	.06
le Breton	0	0
la semaine	.88	.81
quelle semaine	.12	.16
le semestre	.28	.28
quel semestre	0	.02

5. Conclusions

We see the prospects for this approach to UR learning as very bright.

- Distributions over URs that allow lexical patterns to be captured can be learned alongside restrictive grammars that encode a language's generalizations.

This treatment of hidden structure (not just URs) is similarly part of ongoing work on stress learning:

- Learning stress constraints with hidden (foot) structure.
- Learning syllable weight and stress simultaneously.

References

- Apoussidou, Diana. 2007. *The Learnability of Metrical Phonology*. PhD dissertation. University of Amsterdam.
- Boersma, Paul. 1999. Phonology-semantics interaction in OT, and its acquisition. In Robert Kirchner, Wolf Wikeley & Joe Pater (eds.), *Papers in Experimental and Theoretical Linguistics 6*: 24-35. Edmonton: University of Alberta.
- Coetzee, Andries & Joe Pater. to appear. The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.), *Handbook of Phonological Theory, 2nd edition*. [ROA-946].
- Dell, François. 1973. *Les règles et les sons*. Orléans: Imprimerie Nouvelle. [trans. 1980 by Catherine Cullen as *Generative Phonology and French Phonology*. Cambridge University Press.]
- Eisenstat, Sarah. 2009. *Learning Underlying Forms with MaxEnt*. MA thesis. Brown University.
- Eychenne, Julien. 2006. *Aspects de la phonologie du schwa dans le français contemporain optimalité, visibilité prosodique, gradience*. PhD dissertation. Université de Toulouse-Le Mirail.

Eychenne, Julien & Elissa Pustka. 2007. The initial position in Southern French: elision, suppletion, emergence. In Jean-Pierre Angoujard & Olivier Crouzet (eds.), *Proceedings of JEL'2007*, 199-204. Université de Nantes.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson & Ö. Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111-120. Stockholm: Stockholm University.

Hansen, A. 1994. Étude du e caduc—stabilisation en cours et variations lexicales. *Journal of French Language Studies* 4: 25-54.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*, 158-203. Cambridge: Cambridge University Press. [ROA-327].

Inkelas, Sharon, C. Orhan Orgun & Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of the grammar. In Iggy Roca (ed.), *Constraints and Derivations in Phonology*, 393-418. Oxford: Clarendon Press.

Jarosz, Gaja. 2006. *Rich Lexicons and Restrictive Grammars – Maximum Likelihood Learning in Optimality Theory*. PhD dissertation. Johns Hopkins University.

Jesney, Karen & Anne-Michelle Tessier. to appear. Biases in Harmonic Grammar: the road to restrictive learning. *Natural Language and Linguistic Theory*.

Kager, René. 2009. Lexical irregularity and the typology of contrast. In Kristin Hanson & Sharon Inkelas (eds.), *The Nature of the Word: Essays in Honor of Paul Kiparsky*. Cambridge, MA: MIT Press.

Pater, Joe. 2008. Lexically-conditioned variation in Harmonic Grammar. Paper presented at OCP-5. Université de Toulouse-Le Mirail.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (eds.), *Frequency Effects and the Emergence of Lexical Structure*, 137-157. Amsterdam: John Benjamins.

Pierrehumbert, Janet. 2002. Word-specific phonetics . In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory Phonology VII*, 101-139. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46(2-3): 115-154.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell. [ROA-537].

Prince, Alan & Bruce Tesar. 2004. Learning phonotactic distributions. In René

Kager, Joe Pater & Wim Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*, 245-291. Cambridge: Cambridge University Press. [ROA-353].

Racine, Isabelle. 2007. Effacement du schwa dans des mots lexicaux: constitution d'une base de données et analyse comparative. In Jean-Pierre Angoujard & Olivier Crouzet (eds.), *Proceedings of JEL'2007*, 125-130. Université de Nantes.

Racine, Isabelle & François Grosjean. 2002. La production du E caduc facultatif est-elle prévisible? Un début de réponse. *Journal of French Language Studies* 12: 307-326.

Smolensky, Paul. 1996. On the comprehension / production dilemma in child language. *Linguistic Inquiry* 21: 720-731.

Tesar, Bruce. 2006. Faithful contrastive features in learning. *Cognitive Science* 30(5): 863-903.

Walker, Douglas C. 1996. The new stability of unstable -e in French. *French Language Studies* 6: 211-229.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* 30: 945-982.

Weights learned in French schwa deletion simulation:

*ADJUNCT	3.26
MAX	2.57
/s'maine/	1.60
/semestre/	0.99
ALIGN	0.19
*SCHWA	0.00
/semaine/	0.00
/s'mestre/	0.00

*ADJUNCT violation: /VC.CəCV/ → [VC.<C>CV]

ALIGN violation: /V.CəCV/ → [VC.CV]

Test of typology:

Lexicon of 4 suffixes and 4 roots.

Roots		Suffixes	
r1	ra, 'ra	s1	sa, 'sa
r2	re, re:	s2	se, se:
r3	ro, 'ro:	s3	so, 'so:
r4	'ru:	s4	'su:

6 alternating forms with 2 URs each → 12 UR constraints.

10,000 random grammars created by assigning random weights.

Candidates chosen by these grammars in Harmonic Grammar used as target languages.

Learner settings: initial weights = 1.0; $\sigma_2 = 48.0$; $\lambda = 0.3$

The candidate with the highest probability is taken as the winner, regardless of its probability.

Percentage of fully correct languages: 98.61%

Percentage of forms generated correctly per language:

Mean = 99.91%

Median = 100.00%

Standard Deviation = 0.79%

The patterns learned here are categorical.

Mean difference between #1 and #2 probability output: 0.89

(Target: 1.00)

Mean entropy of output probability distribution: 4.61 bits

(Target: 0.00 bits. Maximum entropy of data: 39.81 bits)