how to make EM do what you want

kevin knight

simple substitution cipher

- Ciphertext: **`NOEEI TIMEO ...`**
- Plaintext: **`HELLO WORLD ...`**

- Encipherment table is 1-to-1 both ways:

```
PLAIN:   ABCDEFGHIJKLMNOPQRSTUVWXYZ
CIPHER:  XYZLOHANBCDEFGIJKMPQRSTUVW
```

- Table is unknown to code-breaker
- What table, if applied to the ciphertext, would yield sensible plaintext?

cryptanalysis and machine translation

- Cryptology was dominated by linguists until World War II, when mathematicians entered the field
- Humans are extremely versatile at cracking ciphers
  - breaking a cipher requires an excellent plaintext “language model”
  - robust to encipherment errors, spelling errors
- Mathematical techniques in cryptology are interesting
  - A. M. Turing’s wartime cryptography work
  - *Statistical Methods in Cryptanalysis* (S. Kullback)
  - *A Mathematical Theory of Cryptography* (C. Shannon)
  - Warren Weaver and MT

KDCY LQZKTLJQX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

KDCY LQZKTLJQX CY MDBCYJQL: "TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC."

A	
B	3
C	8
D	7
E	1
F	3
G	
H	3
I	1
J	3
K	9
L	10
M	6
N	1
O	
P	1
Q	11
R	3
S	
T	7
U	
V	
W	1
X	5
Y	7
Z	2

. . . .

KDCY LQZKTLJQX CY MDBCYJQL: "TR

. . . . . .

HYD FKXC, FQ MKX RLQQIQ HYDL

. . . .

MKL DXCTW RDCDLQ JQMNKXTMB

. . . . .

PTBMYEQL K FKH CY LQZKTL TC."

A		
B	3	
C	8	
D	7	
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	9	
L	10	
M	6	
N	1	.
O		
P	1	.
Q	11	
R	3	.
S		
T	7	
U		
V		
W	1	.
X	5	
Y	7	
Z	2	.

```
        .   .                 .       .
KDCY LQZKTLJQX CY MDBCYJQL: "TR

.   .       .       .   .   .
HYD FKXC, FQ MKX RLQQIQ HYDL

        . .       .   .
MKL DXCTW RDCDLQ JQMNKXTMB

.   .       . .       .
PTBMYEQL K FKH CY LQZKTL TC."
```

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

a .a . . .

KDCY LQZKTLJQX CY MDBCYJQL: “TR

. .a . a . . .

HYD FKXC, FQ MKX RLQQIQ HYDL

a . . . .a

MKL DXCTW RDCDLQ JQMNKXTMB

. . a .a. .a

PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

```
a      e.a  .e              .e     .
KDCY LQZKTLJQX CY MDBCYJQL: "TR
.    .a     .e  a  . ee.e .
HYD FKXC, FQ MKX RLQQIQ HYDL
 a        . .    e .e .a
MKL DXCTW RDCDLQ JQMNKXTMB
.     .e  a .a.      e.a
PTBMYEQL K FKH CY LQZKTL TC."
```

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

didn't create "ae"

a e.ao .e .e o.

KDCY LQZKTLJQX CY MDBCYJQL: "TR

. .a .e a . ee.e .

HYD FKXC, FQ MKX RLQQIQ HYDL

a o. . e .e .a o

MKL DXCTW RDCDLQ JQMNKXTMB

.o .e a .a. e.ao o

PTBMYEQL K FKH CY LQZKTL TC."

don't like "ao" – back up!

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

a o e.a .e o o.e .

KDCY LQZKTLJQX CY MDBCYJQL: “TR

.o .a .e a . ee.e .o

HYD FKXC, FQ MKX RLQQIQ HYDL

a . . e .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.e a .a. o e.a

PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 6  #### V
Z 2  .
```

a o re.a r.e o o.e f

KDCY LQZKTLJQX CY MDBCYJQL: “TR

.o .a .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar . f re .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.er a .a. o re.a r

PTBMYEQL K FKH CY LQZKTL TC.”

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 6  #### V
Z 2  .
```

a o re.a r.e o o.e f

KDCY LQZKTLJQX CY MDBCYJQL: "TR

.o .a .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar . f re .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.er a .a. o re.a r

PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ C D T M ~~Y~~ X
frequent English letters: ~~e~~ t ~~o~~ ~~a~~ n i ~~r~~ s h

A		
B	3	
C	8	
D	7	#
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	9	##### V
L	10	##
M	6	#
N	1	.
O		
P	1	.
Q	11	######## V
R	3	.
S		
T	7	### V
U		
V		
W	1	.
X	5	
Y	6	#### V
Z	2	.

a no re.air.e no no.e if

KDCY LQZKTLJQX CY MDBCYJQL: "TR

.o .a n .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ni. f n re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. no re.air in

PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ C D T M ~~Y~~ X
frequent English letters: ~~e~~ t ~~o~~ ~~a~~ n i ~~r~~ s h

A		
B	3	
C	8	
D	7	#
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	9	##### V
L	10	##
M	6	#
N	1	.
O		
P	1	.
Q	11	######## V
R	3	.
S		
T	7	### V
U		
V		
W	1	.
X	5	
Y	6	#### V
Z	2	.

a to re.air.e to to.e if

KDCY LQZKTLJQX CY MDBCYJQL: "TR

.o .a t .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ti. f t re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. to re.air it

PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ ~~C~~ D ~~T~~ M ~~Y~~ X
frequent English letters: ~~e~~ ~~t~~ ~~o~~ ~~a~~ n ~~i~~ ~~r~~ s h

A		
B	3	
C	8	
D	7	#
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	9	##### V
L	10	##
M	6	#
N	1	.
O		
P	1	.
Q	11	######## V
R	3	.
S		
T	7	### V
U		
V		
W	1	.
X	5	
Y	6	#### V
Z	2	.

a to repair.e to to.e if

KDCY LQZKTLJQX CY MDBCYJQL: "TR

.o .a t .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ti. f t re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. to repair it

PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: Q L K C D T M Y X
frequent English letters: e t o a n i r s h

A		
B	3	
C	8	
D	7	#
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	9	##### V
L	10	##
M	6	#
N	1	.
O		
P	1	.
Q	11	######## V
R	3	.
S		
T	7	### V
U		
V		
W	1	.
X	5	
Y	6	#### V
Z	2	.

auto repairmen to customer if

KDCY LQZKTLJQX CY MDBCYJQL: "TR

you wait we can freeze your

HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics

MKL DXCTW RDCDLQ JQMNKXTMB

discover a way to repair it

PTBMYEQL K FKH CY LQZKTL TC."

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 9  ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 11 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 6  #### V
Z 2  .
```

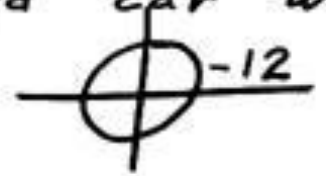
Zodiac killer cipher (unsolved)

340 characters (64 unique)
ASCII version here

```
HER>pl^VPk|?LTG~d
Np+B$4O7DWY.<8Kf=
By}]M+UZGW$=L4@HJ
Spp3^l18VMpO++RK~
52M+@!td|*FP+gnk/
p1R^FlO-8d[kF>~D$
4*+Kq7{~U]XGV.@L|
$G~Jft4O+5NY@+6L2
d<M+b+ZR~FB]yAinK
-@lUV+^J+Op3<FBy-
U+R/*!E|DYBpbTMKO
~<]lRJ|8*T*M.+gBF
@i2Sy4+NI*FB]${1R
lGFN^f*~*b.]V*!++
yBX?8}n2[E>VUZ*-+
|].m@BK$Op^.fMqG~
R]T+L*~[<+FlWB|=L
++=WC@W]POSHT/$=p
|FkdW<3!D5YOB8-[]
>MDHNpkS@ZO1A|K{+
```

This is the Zodiac speaking

I have become very upset with the people of San Fran Bay Area. They have <u>not</u> complied with my wishes for them to wear some nice ⊕ buttons. I promised to punish them if they didnot comply, by anilating a full School Buss. But now school is out for the summer, so I punished them in an another way. I shot a man sitting in a parked car with a .38.



SFPD-0

The Map coupled with this code will tell you where the bomb is set. You have untill next Fall to dig it up. ⊕

C Δ J I ■ O ϰ ⊥ A M ⊐ ▲ Ω O R T G
X ⊙ F D V ⊂ ◪ H C E L ⊕ P W Δ

another Zodiac cipher (this one solved)

408 characters,
54 unique

(first of three parts)

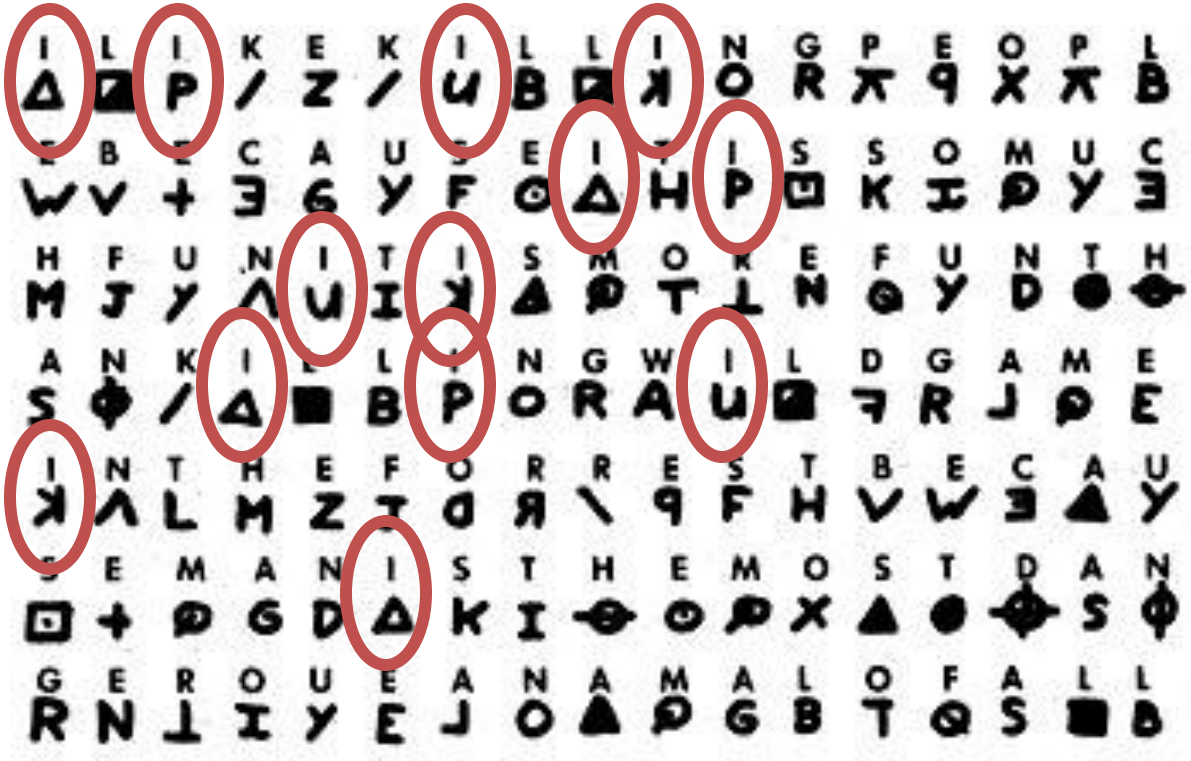
another Zodiac cipher (this one solved)

408 characters,
54 unique

(first of three parts)

note spelling errors in plaintext

another Zodiac cipher (this one solved)



(first of three parts)

408 characters,
54 unique

HOMOPHONIC CIPHER

Frequent plaintext letters have multiple ciphertext renderings.

Ciphertext letter distribution is therefore uniform – harder to crack.

Nondeterministic in the *enciphering* direction. Deterministic in the *deciphering* direction.

another Zodiac cipher (this one solved)

(first of three parts)

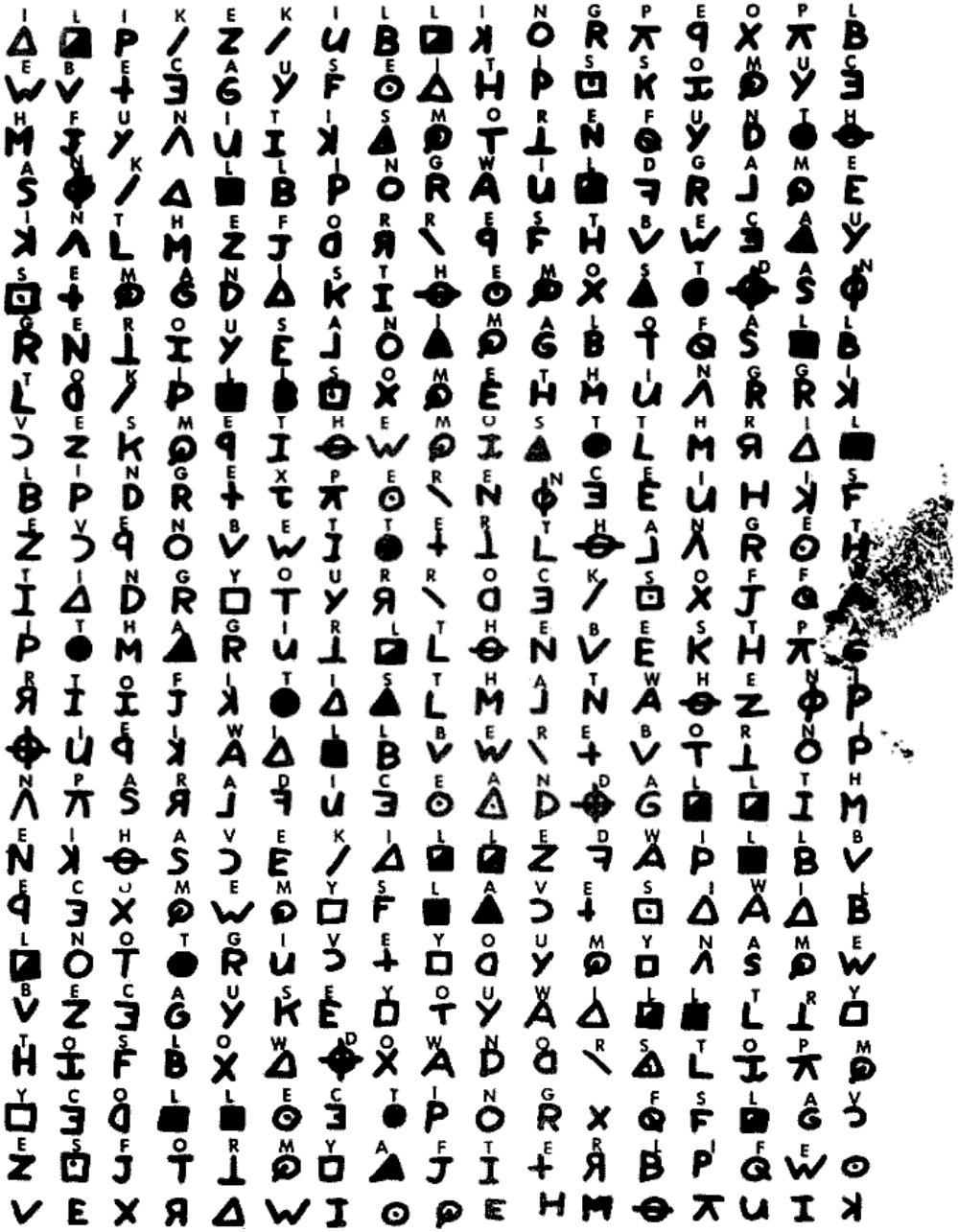
408 characters,
54 unique

HOMOPHONIC CIPHER

Frequent plaintext letters have multiple ciphertext renderings.

Ciphertext letter distribution is therefore uniform – harder to crack.

Nondeterministic in the *enciphering* direction. Deterministic in the *deciphering* direction.



last 18 characters seem to be nonsense

Solution:

PT	CT
A	1 3 G l S
B	V
C	e
D	@ f
E	+ E I N P W Z
F	J Q
G	R
H	= M
I	2 k P U
J	
K	/
L	4 7 B
M	q
N	^ $ D O
O	; d T X
P	:
Q	
R	! \ r
S	1 3 6 F K
T	* H I L N
U	Y
V	]
W	A 2?
X	t
Y	5
Z	

last 18 characters seem to be nonsense

Solution:

PT	CT
A	1 3 G l S
B	V
C	e
D	@ f
E	+ E I N P W Z
F	J Q
G	R
H	= M
I	2 k P U
J	
K	/
L	4 7 B
M	q
N	^ $ D O
O	; d T X
P	:
Q	
R	! \ r
S	1 3 6 F K
T	* H I L N
U	Y
V	]
W	A 2?
X	t
Y	5
Z	

name transliteration

(or, a not-so-secret Japanese code)

- Ciphertext: アンジラナイト
  a n ji ra na i to
- Plaintext: Angela Knight

W. WEAVER

"When I look at Japanese katakana, I say to myself, this is really written in English… I will now proceed to decode."

Attacked first in [Knight & Graehl, 97].
Need to develop a table of sound translation patterns between Japanese and English.

E.g., English T → Japanese {T, TO, …}
English L → Japanese {R, RU, …}
English IH → Japanese {I, II, E, EE, …}

With such a table, we can statistically decode new names.

English/Japanese sound correspondences learned by EM.

[Knight & Graehl 97]

Note multiple mappings in both directions.

e	j	$P(j \mid e)$
AA	o	0.566
	a	0.382
	a a	0.024
	o o	0.018
AE	a	0.942
	y a	0.046
AH	a	0.486
	o	0.169
	e	0.134
	i	0.111
	u	0.076
AO	o	0.671
	o o	0.257
	a	0.047
AW	a u	0.830
	a w	0.095
	o o	0.027
	a o	0.020
	a	0.014
AY	a i	0.864
	i	0.073
	a	0.018
	a i y	0.018
B	b	0.802
	b u	0.185
CH	ch y	0.277
	ch	0.240
	tch i	0.199
	ch i	0.159
	tch	0.038
	ch y u	0.021
	tch y	0.020
D	d	0.535
	d o	0.329
	dd o	0.053
	j	0.032
DH	z	0.670
	z u	0.125
	j	0.125
	a z	0.080
EH	e	0.901
	a	0.069
ER	a a	0.719
	a	0.081
	a r	0.063
	e r	0.042
	o r	0.029

e	j	$P(j \mid e)$
EY	e e	0.641
	a	0.122
	e	0.114
	e i	0.080
	a i	0.014
F	h	0.623
	h u	0.331
	hh	0.019
	a h u	0.010
G	g	0.598
	g u	0.304
	gg u	0.059
	gg	0.010
HH	h	0.959
	w	0.014
IH	i	0.908
	e	0.071
IY	i i	0.573
	i	0.317
	e	0.074
	e e	0.016
JH	j	0.329
	j y	0.328
	j i	0.129
	jj i	0.066
	e j i	0.057
	z	0.032
	g	0.018
	jj	0.012
	e	0.012
K	k	0.528
	k u	0.238
	kk u	0.150
	kk	0.043
	k i	0.015
	k y	0.012
L	r	0.621
	r u	0.362
M	m	0.653
	m u	0.207
	n	0.123
	n m	0.011
N	n	0.978
NG	n g u	0.743
	n	0.220
	n g	0.023

e	j	$P(j \mid e)$
OW	o	0.516
	o o	0.456
	o u	0.011
OY	o i	0.828
	o o i	0.057
	i	0.029
	o i y	0.029
	o	0.027
	o o y	0.014
	o o	0.014
P	p	0.649
	p u	0.218
	pp u	0.085
	pp	0.045
PAUSE	pause	1.000
R	r	0.661
	a	0.170
	o	0.076
	r u	0.042
	u r	0.016
	a r	0.012
S	s u	0.539
	s	0.269
	sh	0.109
	u	0.028
	ss	0.014
SH	sh y	0.475
	sh	0.175
	ssh y u	0.166
	ssh y	0.088
	sh i	0.029
	ssh	0.027
	sh y u	0.015
T	t	0.463
	t o	0.305
	tt o	0.103
	ch	0.043
	tt	0.021
	ts	0.020
	ts u	0.011
TH	s u	0.418
	s	0.303
	sh	0.130
	ch	0.038
	t	0.029

e	j	$P(j \mid e)$
UH	u	0.794
	u u	0.098
	dd	0.034
	a	0.030
	o	0.026
UW	u u	0.550
	u	0.302
	y u u	0.109
	y u	0.021
V	b	0.810
	b u	0.150
	w	0.015
W	w	0.693
	u	0.194
	o	0.039
	i	0.027
	a	0.015
	e	0.012
Y	y	0.652
	i	0.220
	y u	0.050
	u	0.048
	b	0.016
Z	z	0.296
	z u	0.283
	j	0.107
	s u	0.103
	u	0.073
	a	0.036
	o	0.018
	s	0.015
	n	0.013
	i	0.011
	sh	0.011
ZH	j y	0.324
	sh i	0.270
	j i	0.173
	j	0.135
	a j y u	0.027
	sh y	0.027
	s	0.027
	a j i	0.016

name transliteration

L AE M P (english)
R A N P U (japanese)

S AA K ER
S A K K A A

EH R K R A F T
E A K U R A H U T O

TRAIN

e	j	P(j | e)
AA	e	0.566
	a	0.382
	e a	0.024
	o o	0.018
AE	a	0.942
	y a	0.046
AH	a	0.486
	o	0.169
	e	0.134
	i	0.111
	u	0.076
AO	o	0.471
	o o	0.257
	a	0.047
AW	a u	0.836
	a w	0.095
	o o	0.027
	a o	0.026
	a	0.014
AY	a i	0.964
	i	0.073
	a	0.018
	a i y	0.018
B	b	0.802
	b u	0.185
CH	ch y	0.277
	ch	0.240
	tch i	0.190
	ch i	0.159
	tch	0.038
	ch y u	0.021
	tch y	0.020
D	d	0.535
	d o	0.229
	dd o	0.053
	j	0.032
DH	z	0.470
	z u	0.125
	j	0.125
	z z	0.080
EH	e	0.901
	a	0.069
ER	a a	0.716
	a	0.081
	a r	0.063
	a r	0.042
	e r	0.020

e	j	P(j | e)
EY	e e	0.641
	a	0.122
	e	0.114
	e i	0.080
	a i	0.014
F	h	0.623
	h u	0.331
	hh	0.019
	a h u	0.016
G	g	0.598
	g u	0.304
	gg u	0.058
	gg	0.016
HH	h	0.959
	e	0.014
IH	i	0.908
	e	0.071
IY	i i	0.372
	i	0.317
	e	0.074
	e e	0.016
JH	j	0.329
	j y	0.328
	j i	0.129
	jj i	0.066
	e j i	0.057
	z	0.032
	g	0.018
	jj	0.012
	a	0.012
K	k	0.528
	k u	0.238
	kk u	0.150
	kk	0.043
	k i	0.015
	k y	0.012
L	r	0.621
	r u	0.362
M	m	0.653
	m u	0.207
	n	0.123
	n m	0.011
N	n	0.976
NG	n g u	0.747
	n	0.226
	n g	0.023

e	j	P(j | e)
OW	o	0.516
	o o	0.456
	o u	0.011
OY	o i	0.828
	o o i	0.057
	i	0.029
	o i y	0.029
	o	0.027
	o o y	0.014
	o o	0.014
P	p	0.649
	p u	0.218
	pp u	0.085
	pp	0.045
PAUSE	pause	1.000
R	r	0.661
	a	0.170
	o	0.076
	r u	0.042
	u r	0.016
	a r	0.012
S	s u	0.539
	s	0.269
	sh	0.169
	u	0.028
	ss	0.014
SH	sh y	0.475
	sh	0.175
	ssh y u	0.166
	ssh y	0.088
	sh i	0.029
	ssh	0.027
	sh y u	0.015
T	t	0.463
	t o	0.305
	tt o	0.183
	ch	0.043
	tt	0.021
	to	0.020
	to u	0.011
TH	s u	0.418
	s	0.383
	sh	0.130
	ch	0.038
	t	0.029

e	j	P(j | e)
UH	u	0.794
	u u	0.088
	dd	0.034
	a	0.030
	o	0.026
UW	u u	0.550
	u	0.302
	y u u	0.109
	y u	0.021
V	b	0.610
	b u	0.156
	w	0.015
W	w	0.693
	u	0.194
	o	0.039
	i	0.027
	a	0.015
	e	0.012
Y	y	0.652
	i	0.226
	y u	0.056
	u	0.048
	b	0.016
Z	z	0.296
	z u	0.283
	j	0.107
	z u	0.103
	u	0.073
	a	0.036
	o	0.018
	e	0.015
	a	0.013
	i	0.011
	sh	0.011
ZH	j y	0.324
	sh i	0.270
	j i	0.173
	j	0.135
	s j y u	0.027
	sh y	0.027
	z	0.027
	s j i	0.016

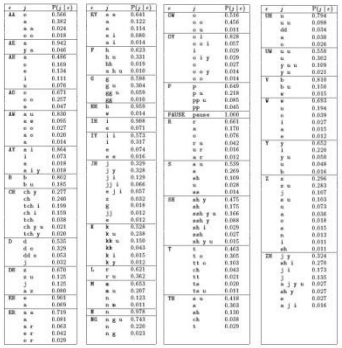
EM method given in
[Knight & Graehl 97]

name transliteration

L AE M P
S AA K ER
EH R K R A F T
…

K U R I S U T I N A
B U R I T A N I
R I N Z I
….

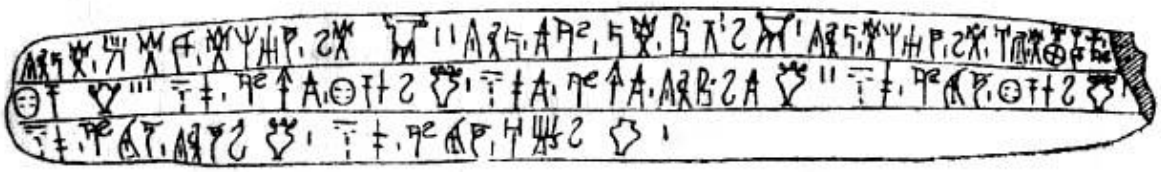
TRAIN



no method yet published for training on non-parallel data

ancient civilizations

- Ciphertext:

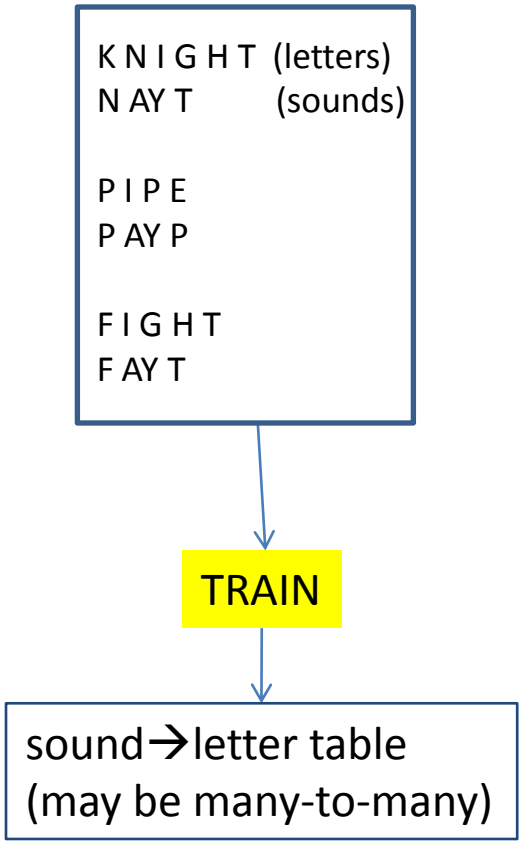


- Plaintext:

  A big vessel with 4 grips, Two big vessels with 3 grips,
  A small vessel with 4 grips, A small vessel with 3 grips, …

- Linear B, Mayan hieroglyphs, Egyptian hieroglyphs, Easter Island glyphs...
- First step is to assign phonetic values to signs
- Essentially: text-to-speech

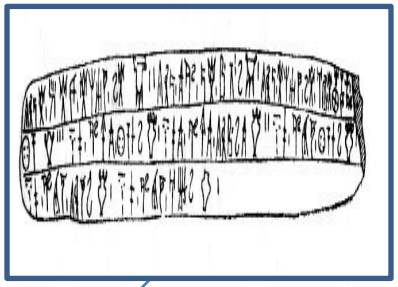
assigning sounds to written signs



basis of text-to-speech systems

assigning sounds to written signs

B UH N AE N UH
L IH Z F EH R
T R IH P L TH UH
P L AE N



TRAIN

sound→letter table
(may be many-to-many)

[Knight & Yamada, 1999;
Knight et al 2006]

speech recognition as decipherment

100,000 hours of recorded speech

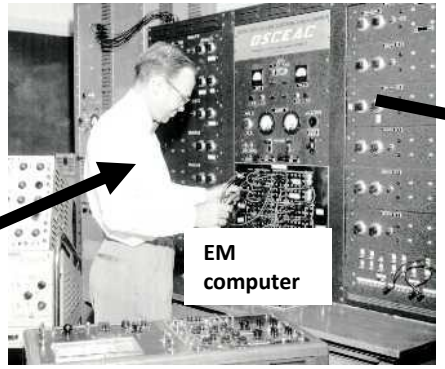
1,000,000,000 words of text

DECIPHER

pronunciation (and other) tables to power a speech recognition engine

deciphering keyboard clicks

Keyboard Acoustic Emanations Revisited
Proceedings of the 12th ACM Conference on Computer and Communications Security, November 2005, pp. 373-382
Zhuang, Zhou, Tygar



APPENDIX

A. RECOVERED TEXT EXAMPLES

Text recognized by the HMM classifier, with cepstrum features (underlined words are wrong),

> the big money fight has drawn the shoporo od dosens of companies in the entertainment industry as well as attorneys gnnerals on states, who fear the fild shading softwate will encourage illegal acyivitt, srem the grosth of small arrists and lead to lost cobs and dimished sales tas revenue.

Text after spell correction using trigram decoding,

> the big money fight has drawn the support of dozens of companies in the entertainment industry as well as attorneys generals in states, who fear the film sharing software will encourage illegal activity, stem the growth of small artists and lead to lost jobs and finished sales tax revenue.

Original text. Notice that it actually contains two typos, one of which is fixed by our spelling corrector.

> the big money fight has drawn the support of dozens of companies in the entertainment industry as well as attorneys gnnerals in states, who fear the file sharing software will encourage illegal activity, stem the growth of small artists and lead to lost jobs and dimished sales tax revenue.

holy grail: machine translation

1,000,000,000
words of English text

1,000,000,000
words of foreign text

DECIPHER

phrase tables to power a statistical MT system

will show some initial results, time permitting

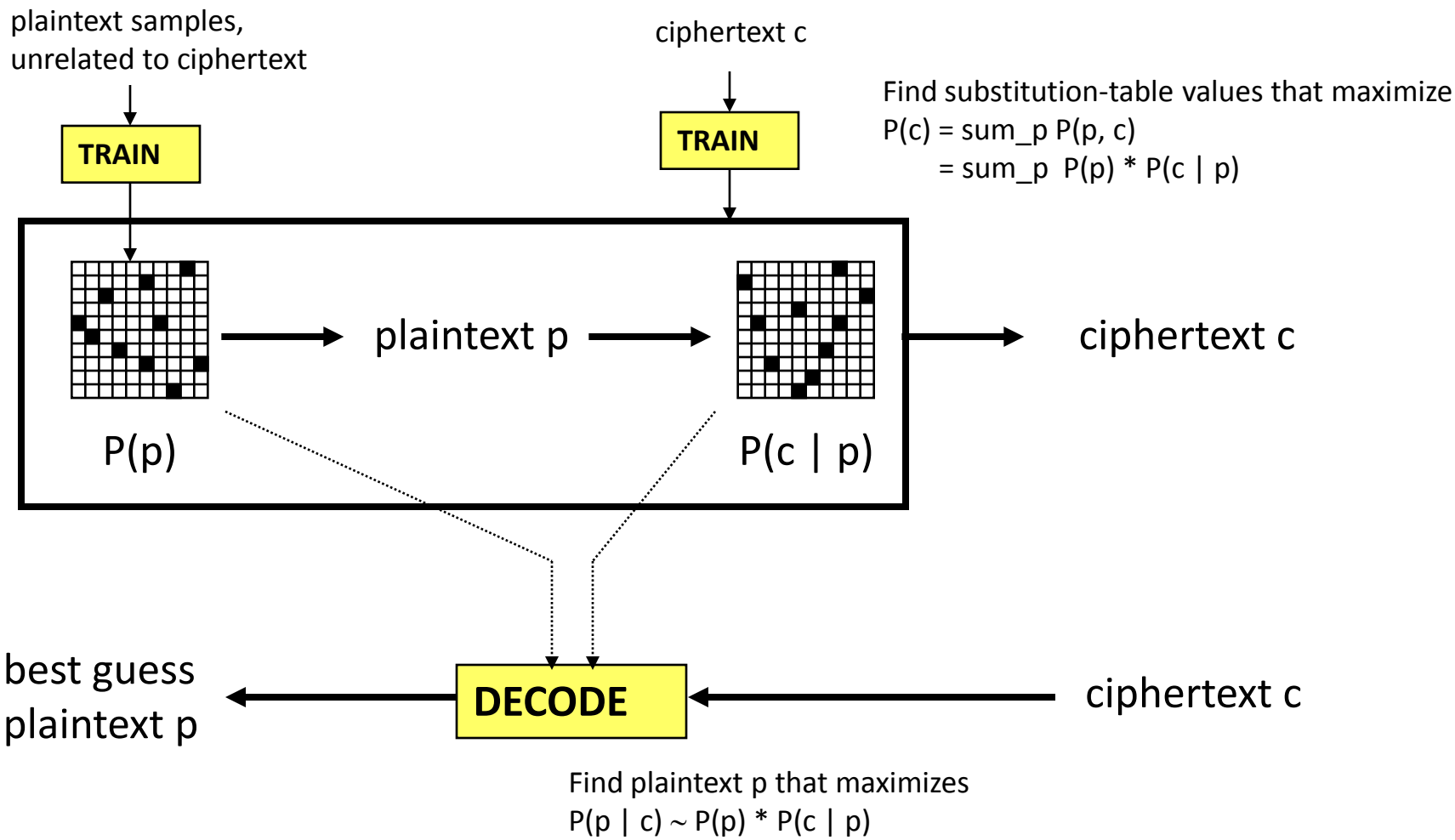
features of problems

	Determinist. decoding	Determinist. encoding	Input & output of same length	Spaces in cipher?	Mono-tone subst?	NULL free	Dictionary of legal sequences
Simple substitution	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
Zodiac cipher	mostly	no	yes	no	yes	?	yes
Archaeology letter-to-sound	no	no	no	no	mostly	no	?
Name translation	no	no	no	no	yes	no	yes
Voynich MS	?	?	?	?	?	?	?
Speech reco	no	no	no	no	yes	no	no
Machine translation	no	no	no	yes	no	no	no

Want to make as few assumptions as possible.
Want to work with as sparse data as possible.
What is the least amount of knowledge required to solve a given problem?

basic technical approach

same approach applies to all problems just discussed

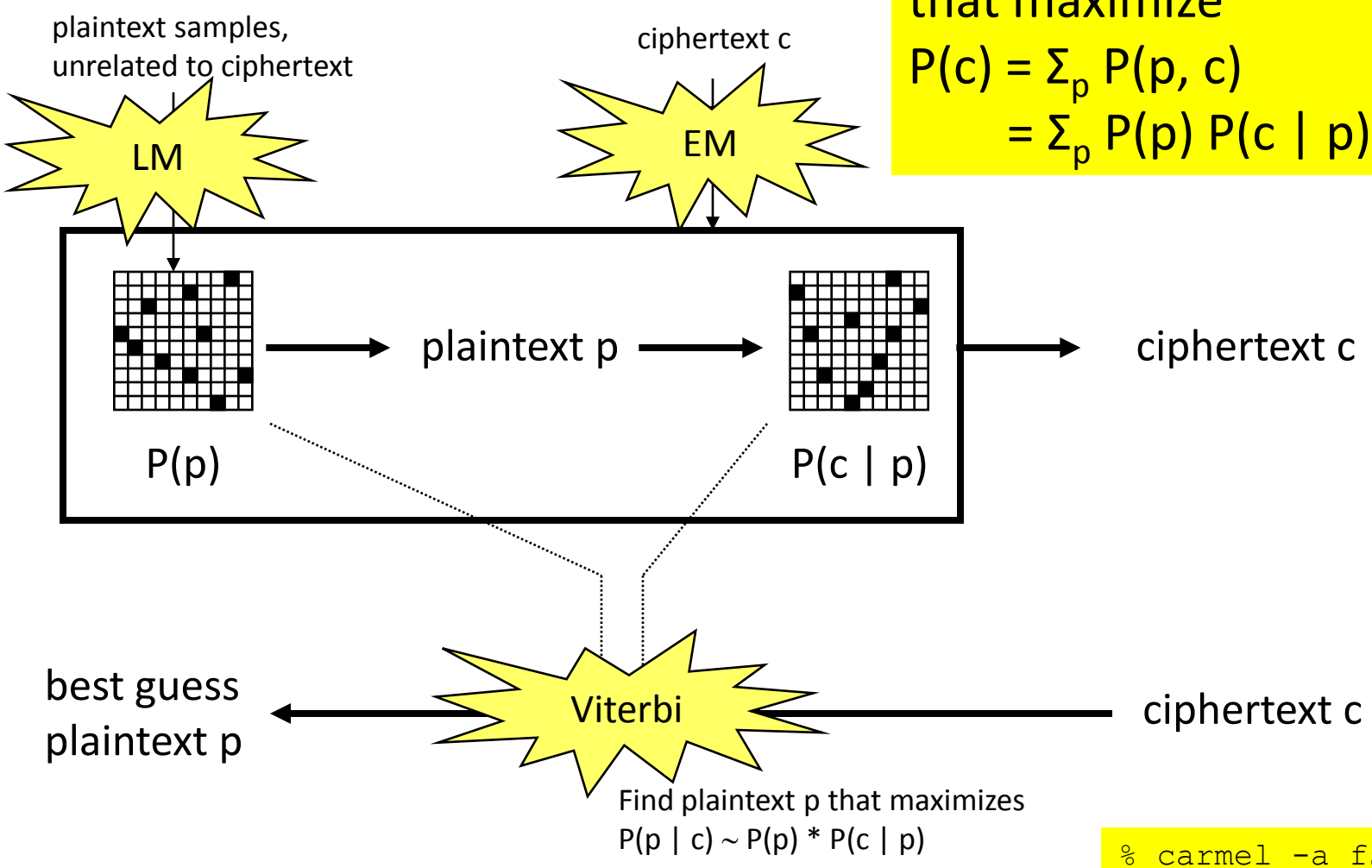


basic technical approach

Find substitution-table values that maximize

$P(c) = \Sigma_p P(p, c)$

$= \Sigma_p P(p) P(c \mid p)$




```
% carmel -a fsa fst >fsa2
% carmel -t fsa2 cipher
```

previous results on decipherment

[knight et al 2006]

- 414-letter substitution cipher
- Attacked with 2-gram and 3-gram <u>letter-based</u> LMs of English
  - Minimal knowledge of English – right up the computer's alley!
- Useful tips:
  - Maximize $P(c) \sim P(e) \cdot P(c \mid e)$ when training
  - Maximize $P(e) \cdot P(c \mid e)^3$ when decoding
  - More LM data helps, important to smooth the LMs
- 2-gram letter-based LM: 10.6% error
- 3-gram letter-based LM: 3.6% error
  - Input: `ingcmpnqsnwf cv fpn owoktvcv hu ihgzsnwfv rqcffnw cw owgcnwf kowazoanv …`
  - Output: `DECIPHERMENT IS THE ANALYSIS OF WOCUMENTS WRITTEN IN ANCIENT LANGUAGES …`

I was satisfied … until my Dad signed Angela up with American Cryptogram Association

Please send relevant contributions, comments, and inquiries to this editor, **QUIPOGAM**:

Aristocrats

Leonard C. Morgan, Jr.
18 Benbrook Circle NE
Roanoke, VA 24012
LmorganJr@aol.com

Dedicated to the Memory of DAMON:

0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0o0

A-1. Fasteners all. K2 [90] (the-4) FLYING DUTCHMAN
CT-AQK AQLKK ILRYHRIFW AQRYPZ AQFA QCWJ HRORWREFARCY
RT-THE THREE PRINCIPAL THINGS THAT HOLD CIVILIZATION
CT-ACPKAQKL FLK AQK ZFMKAD IRY, AQK IFIKL HWRI, FYJ AQK
RT-TOGETHER ARE THE SAFETY PIN THE PAPER CLIP AND THE
CT-ERIIKL.
RT-ZIPPER

A-2. It hurts! K3 [81] REAL NEO
KBQLHJRPZVBP BE P AUQBZBNPNBDV IPBD ERDALHJU NWPN PKKUGNE
NMH NH KHCL IULGUDN HK NWU IHICZPNBHD.

A-3. Fascinating saying. K2 [89] AURION
CT ZIWKDYEH LOESYWC YELZKYGMYFE YL: "Y SF EFM ZFOEM MIU SWTL
RT CHARGING SUNDIAL INSCRIPTION IS: "I DO NOT COUNT THE DAYS
CT FN ZCFOSL WES LIFQUKL; Y FECT ZFOEM MIU LIYEYEH IFOKL!"
RT OR CLOUDS AND SHOWERS I ONLY COUNT THE SHINING HOURS!"

A-4. Cannibal poetry. K2 [91] ANGO-KA
CT NP DAX *JSCN WNVDNLPSCM, YLLW'B WXYNPXW SB "PLD=S=*JSCN."
RT IN THE WARI DICTIONARY FOOD S DEFINED AS NOT A WARI
CT DAXNC WNPPXCB SCX S RLD LY YEP, YLC SRR UED DAX EP=*JSCN
RT THEIR DINNERS ARE A LOT OF FUN FOR ALL BUT THE UN WARI
CT LPX. MEOOM!
RT ONE YUMMY!

A-5. Perspective. K2 [98] CONFUOCO
CZW QFWAQYW *WIYTOBZHQI ZQB BN VWWG Q AWFWAWIUW XNA

head to head, 90-letter cipher

- Cipher:
  - `aqk aqlkk ilryhrifw aqrypz aqfa qcwj hrorwrefacy acpkaqkl flk aqk zfmkad iry aqk ifikl hwri fyj aqk eriikl`
- EM 3-gram LM:
  - `THE THAVE PRONFICAL THINGS THAT BULD FORIVACTION TOGETHER ARE THE WAYESS MIN THE MAKED FLOK AND THE COMPED`

head to head, 90-letter cipher

- Cipher:
  - `aqk aqlkk ilryhrifw aqrypz aqfa qcwj hrorwrefacy acpkaqkl flk aqk zfmkad iry aqk ifikl hwri fyj aqk eriikl`
- EM 3-gram LM:
  - `THE THAVE PRONFICAL THINGS THAT BULD FORIVACTION TOGETHER ARE THE WAYESS MIN THE MAKED FLOK AND THE COMPED`
- Angela:
  - `THE THREE PRINCIPAL THINGS THAT HOLD CIVILIZATION TOGETHER ARE THE SAFETY PIN THE PAPER CLIP AND THE ZIPPER`

Computer has 30% error = 30% of letters are deciphered wrong!

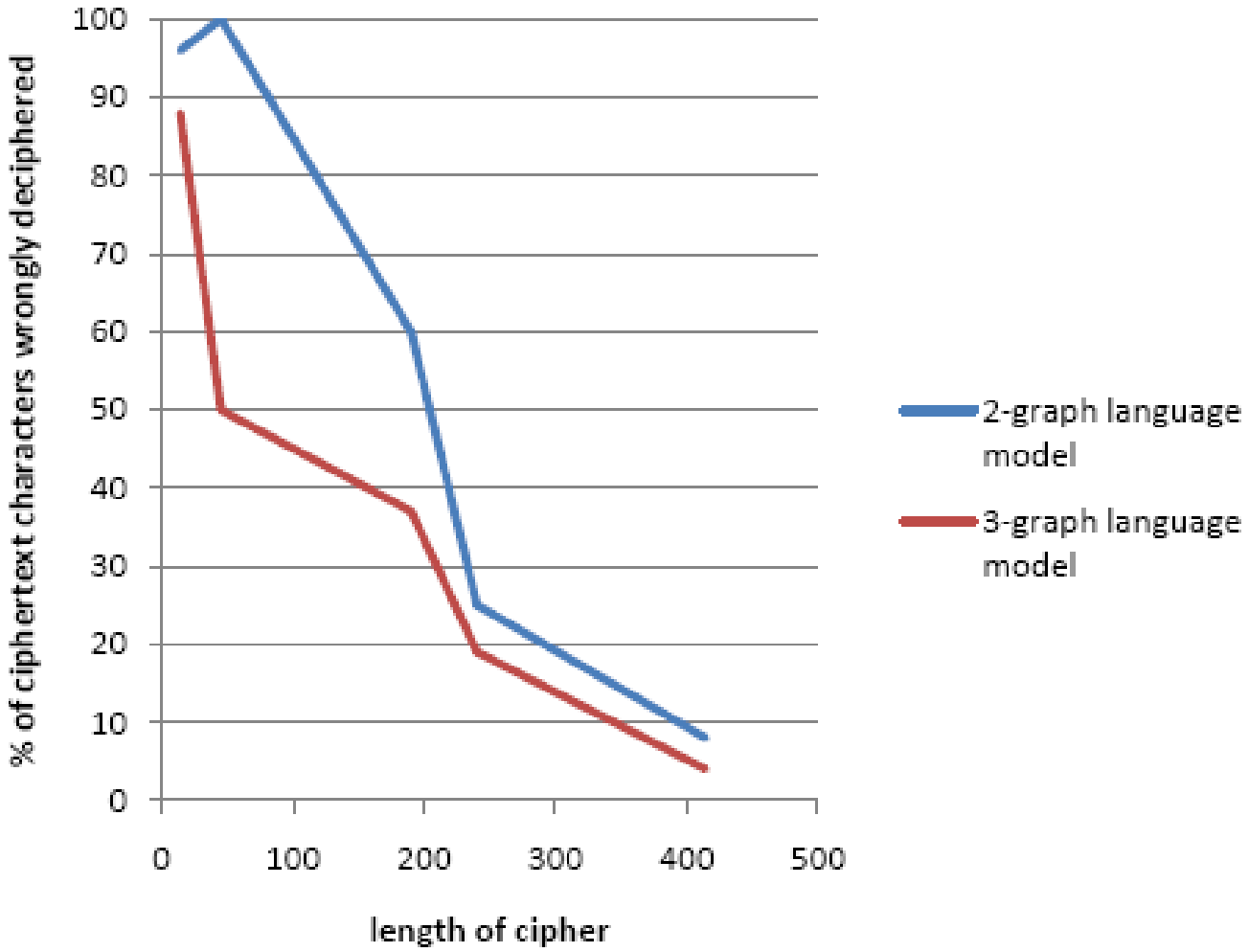
note human robustness to spelling errors in plaintext

hmm, let's try another cipher…

- Cipher:
  - `czw qfwaqyw wiytobzhqi zqb bn vwwg q awfwawiuw xna qicoseocm czqc zw jnetv aqczwa rw janiy czqi awuwic gwcwa huqaczea`
- EM 3-gram LM:
  - `THE PREPACE EXCLUDICAL ING OF WEED A SUREMENCE WAS ANTIONITY THAT OF JABLY GOVIES OF THAND TION SEMENT DEVES CONSTING`

60% error

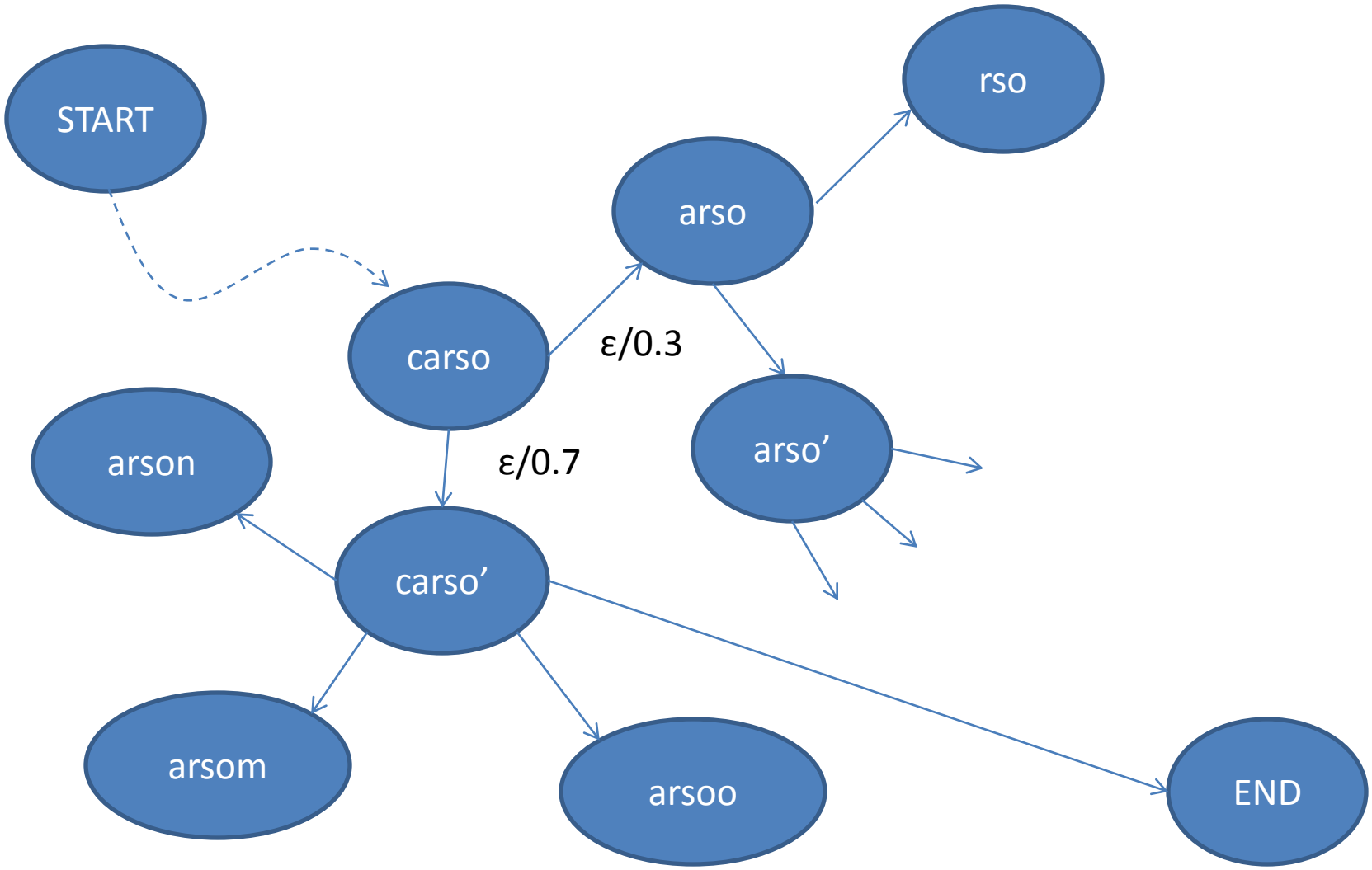
machine solution quality curve



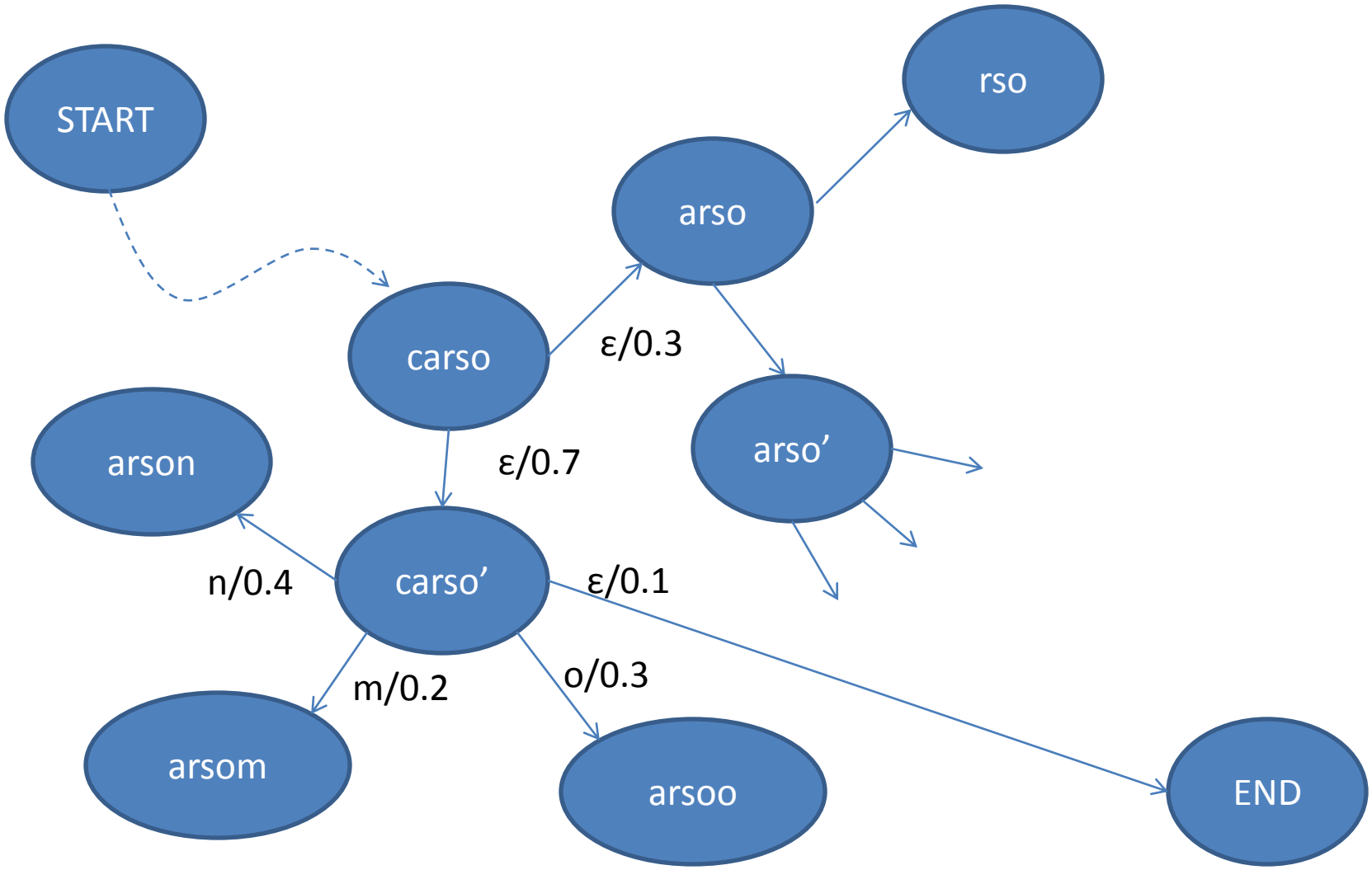
better language models?

- Maybe Angela just knows more English!
- Let’s try 5-gram and 7-gram letter LMs
- Need to build Carmel LMs of arbitrary order

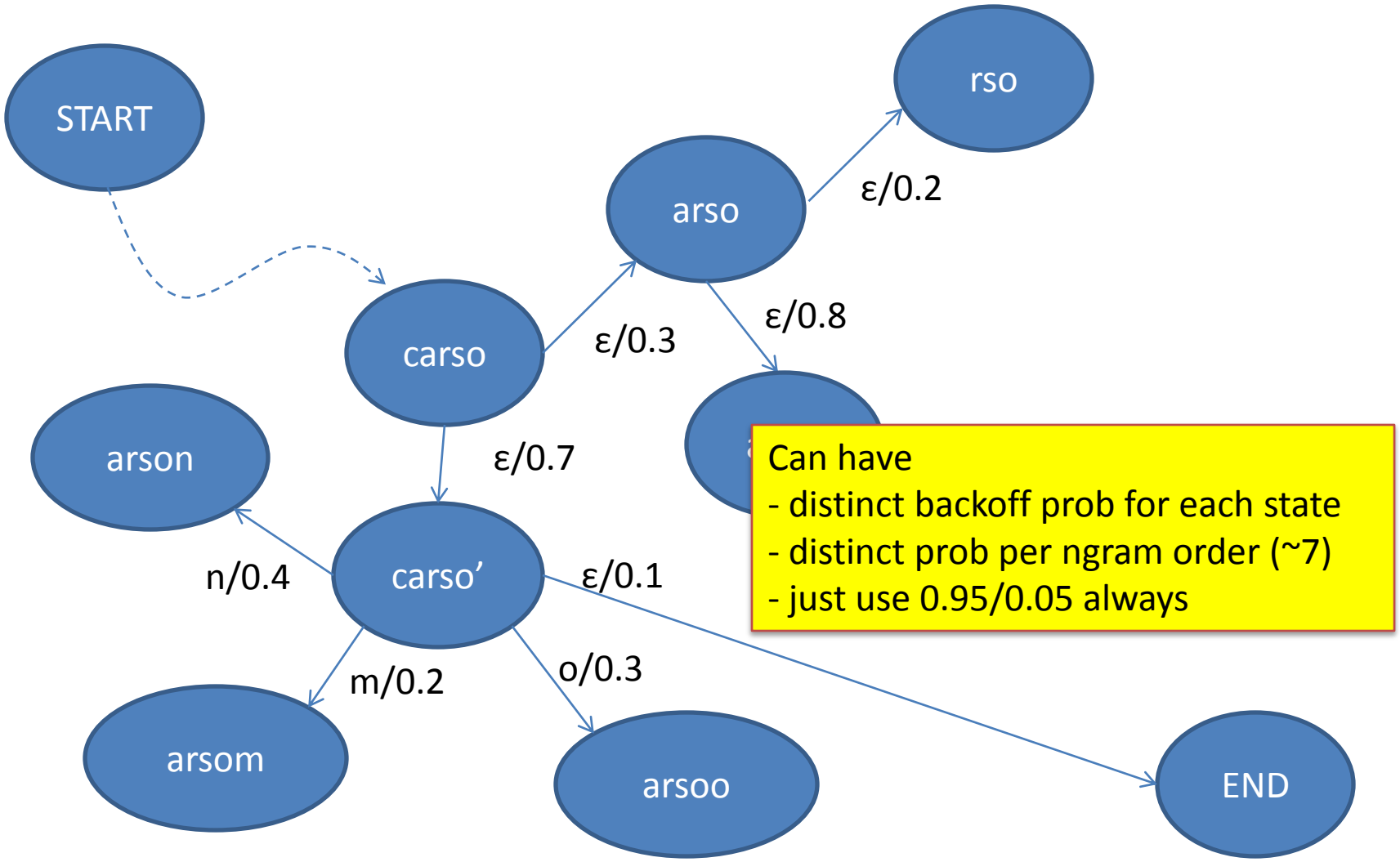
representing LM as a Carmel WFSA



representing LM as a Carmel WFSA



representing LM as a Carmel WFSA



Can have
- distinct backoff prob for each state
- distinct prob per ngram order (~7)
- just use 0.95/0.05 always

code to build smoothed n-gram LM as WFSA

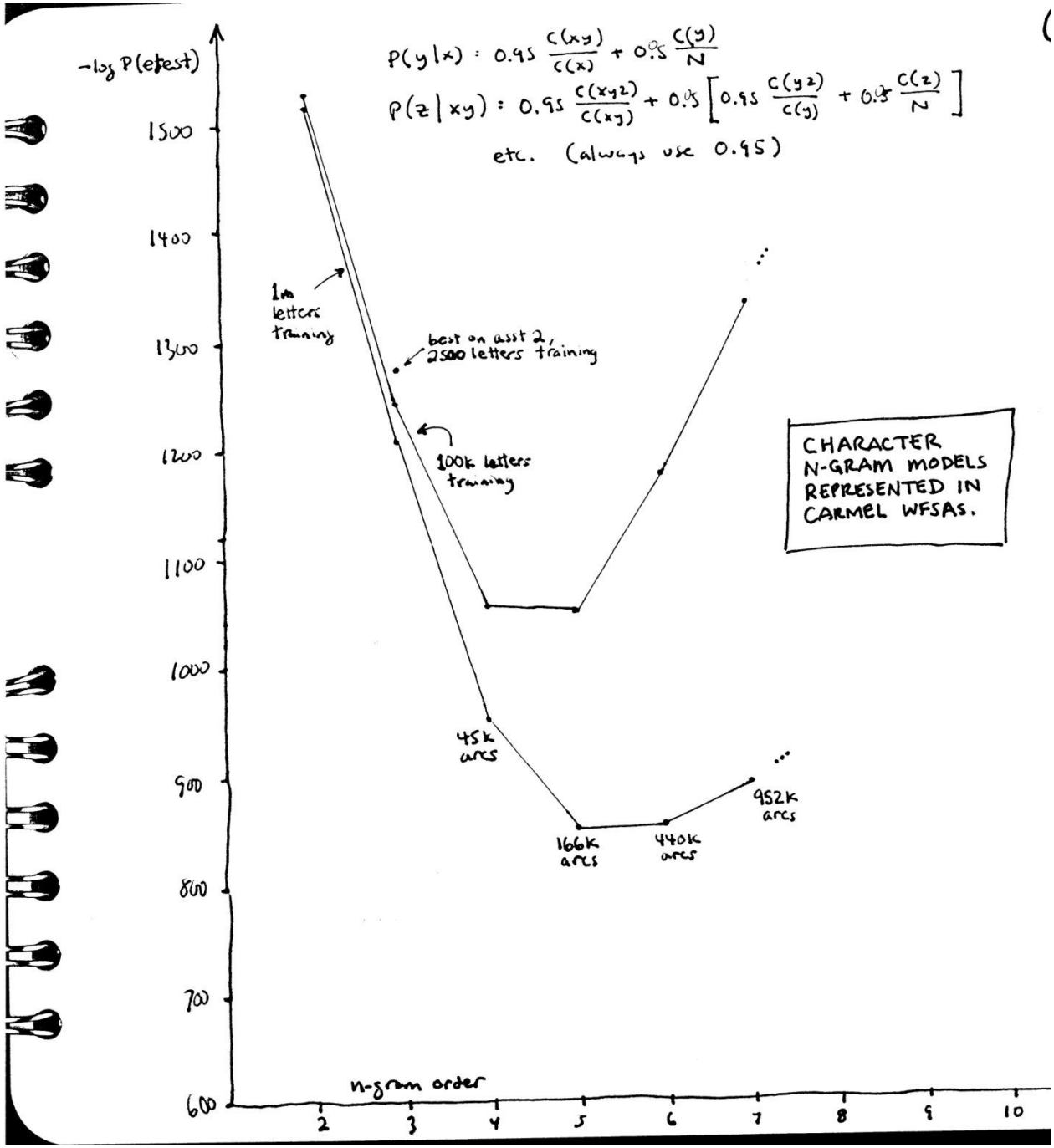
```
CARMEL=/nfs/topaz/graehl/isd/carmel/bin/linux/carmel.static

cat $1 |
head -$2 |
tr 'a-z' 'A-Z' |
tr -d '"' |
gawk 'NF>0' |
sed 's/^/0S0 /' |
sed 's/$/ 0E0/' |
gawk '{for (i=1; i<='$4'; i++)
       for (j=1; j<=(NF-i+1); j++) {
         for (k=j; k<=(j+i-1); k++)
           printf("%s ", $k);
         printf("\n")}}' |
sort -T /tmp -S 2g | uniq -c |
gawk 'BEGIN {printf("0E0-\n(0S0- (000 *e* *e* 1.0))\n")}
    {m = (NF-1);
     if (m==1) back=0.95;
     else if (m==2) back=0.9;
     else if (m==3) back=0.8;
     else if (m==4) back=0.7;
     else if (m==5) back=0.6;
     else back=0.21;
 if (($1 > 1) || (m < '$5')) {
       if ((m==1) && ($2 != "0E0")) {
         printf("(%s- (%s-pr *e* *e* 0.95!))\n",$2,$2);
         printf("(%s- (NULL *e* *e* 0.05!))\n",$2)}
       if ((m==1) && ($2 != "0S0")) {
         if ($2 == "0E0")
           printf("(NULL (%s- *e* *e* %20.10f))\n",$2,
             $1/100000000);
         else
           printf("(NULL (%s- *e* \"%s\" %20.10f))\n",$2,
             $2,$1/100000000)}
       if (m>1) {
         printf("(");
         for (k=2; k<=m; k++) printf("%s-", $k);
         printf("pr (");
         if (m<'$4') printf("%s-",$2);
         for (k=3; k<=(m+1); k++) printf("%s-", $k);
         if ($NF=="0E0")
           printf(" *e* *e* %20.10f))\n",$1/100000000);
         else
           printf(" *e* \"%s\" %20.10f))\n",$NF,$1/100000000)}
```

```
       if ((m>1) && (m<'$4')) {
         printf("(");
         for (k=2; k<=(m+1); k++) printf("%s-", $k);
         printf(" (");
         for (k=2; k<=(m+1); k++) printf("%s-", $k);
         printf("pr *e* *e* %4.2f!))\n",back);
         printf("(");
         for (k=2; k<=(m+1); k++) printf("%s-", $k);
         printf(" (");
         for (k=3; k<=(m+1); k++) printf("%s-", $k);
         printf(" *e* *e* %4.2f!))\n",1.0-back)}}}' |
$CARMEL -sJHn |
sed 's/))/!))/' |
sed 's/0.95!!/0.95!1/' |
sed 's/0.9!!/0.9!2/' |
sed 's/0.8!!/0.8!3/' |
sed 's/0.7!!/0.7!4/' |
sed 's/0.6!!/0.6!5/' |
sed 's/0.21!!/0.2!6/' |
sed 's/0.05!!/0.05!7/' |
sed 's/0.1!!/0.1!8/' |
sed 's/0.2!!/0.2!9/' |
sed 's/0.3!!/0.3!10/' |
sed 's/0.4!!/0.4!11/' |
sed 's/0.79!!/0.8!12/' > zz
$CARMEL -HJtM 10 $3 zz |
$CARMEL -sHJN 0
```

All backoff probs set to 0.95

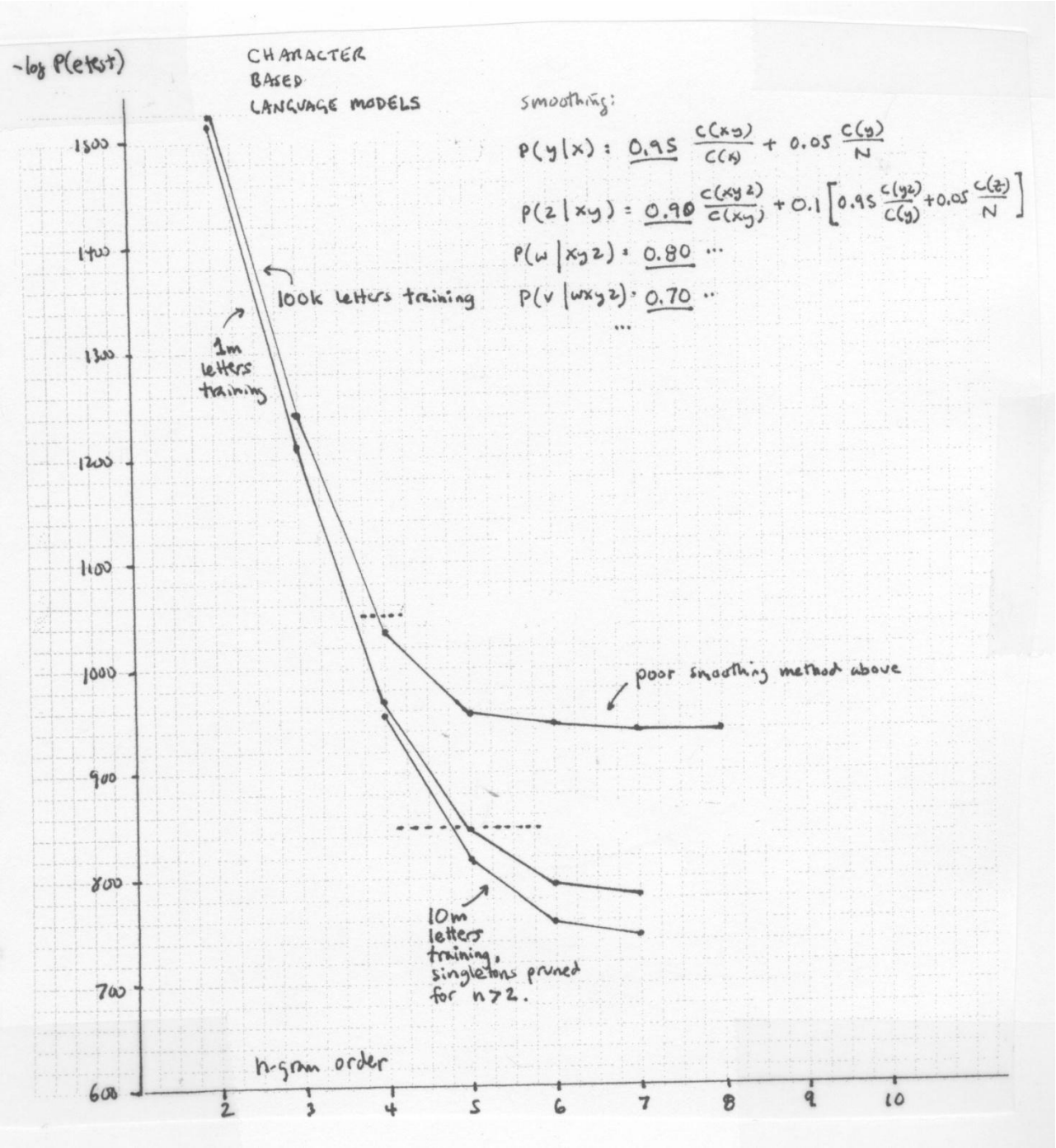
$P(y \mid x) =$
$0.95\ C(xy)/C(x)$
$+\ 0.05\ C(y)/N$

$P(z \mid x\ y) =$
$0.95\ C(xyz)/C(xy)$
$+\ 0.05\ P(z \mid y)$



Manually-set backoff probabilities

0.95 for 2-grams
0.90 for 3-grams
0.80 for 4-grams
0.70 for 5-grams
0.60 for 6-grams
0.20 for 7-grams



stochastic generation

2-gram: … **itariaris s oriorcupunond rke uth** …

3-gram: … **ind thnowelf jusision thad inat of** …

4-gram: … **rece bence on but ther servier** …

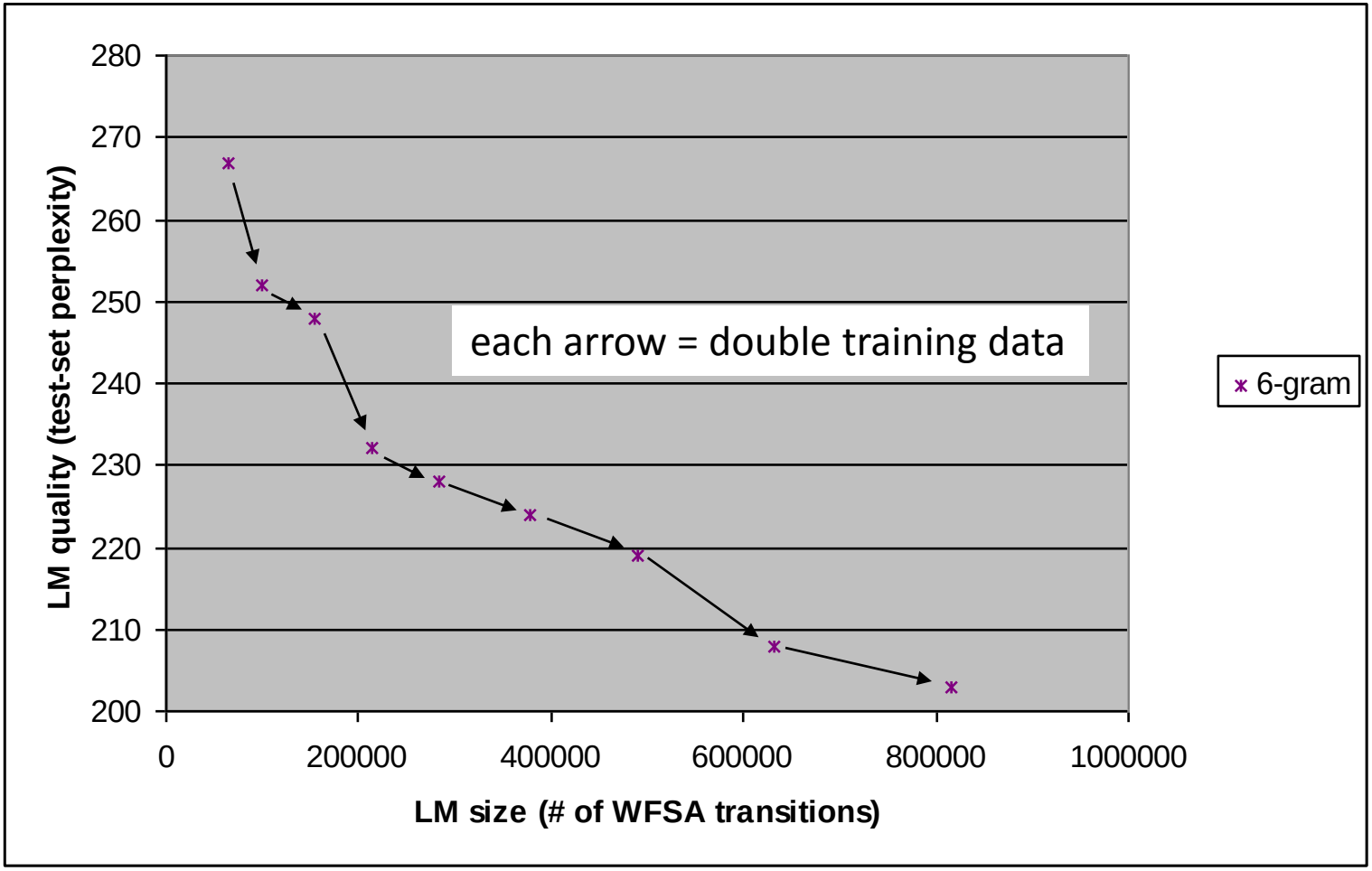
5-gram: … **mrs earned age im on d the perious** …

6-gram: … **a party to possible upon rest of** …

7-gram: … **t our general through approve the** …

```
% carmel -g 1 fsa
```

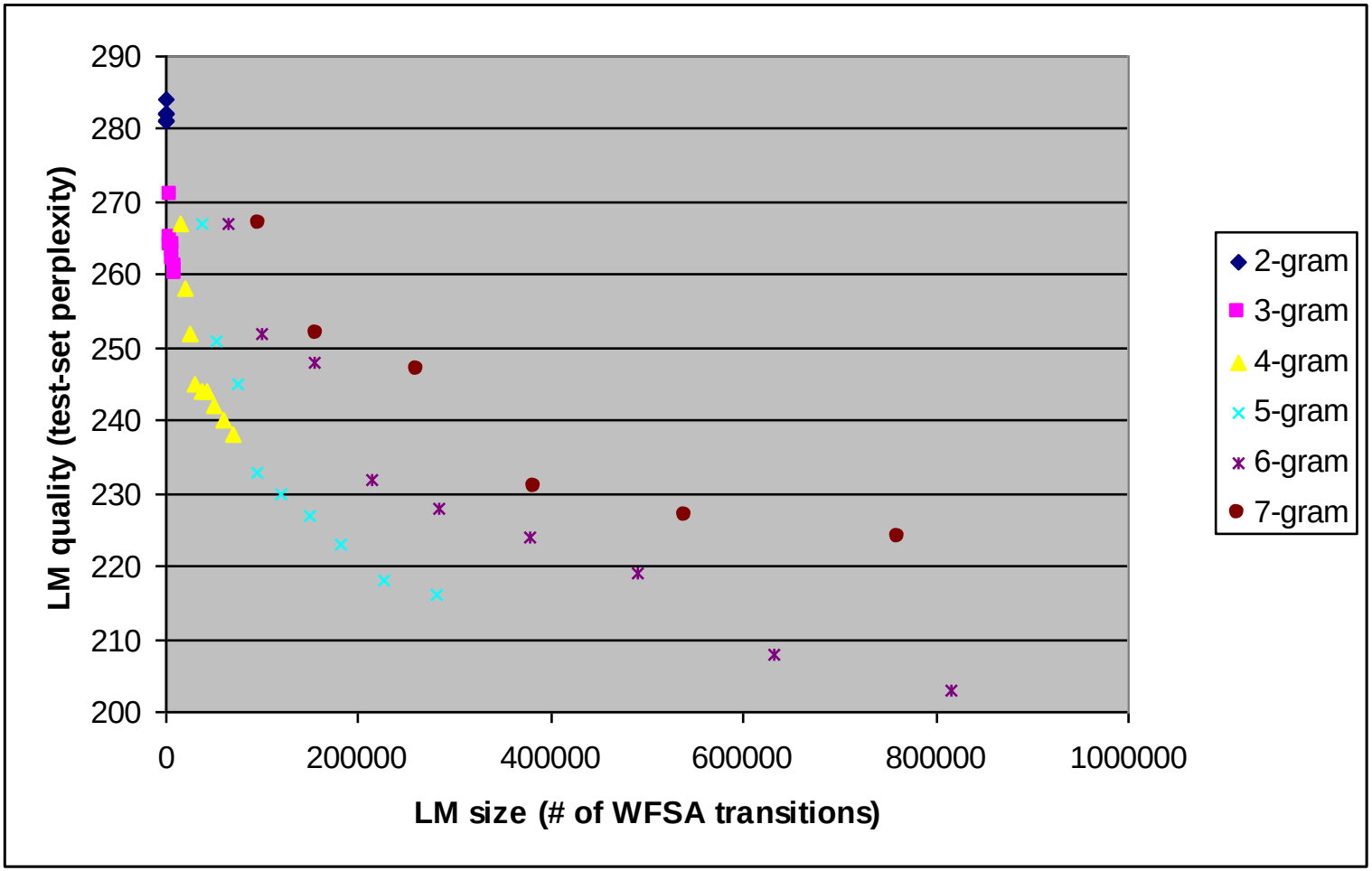
LM memory vs. LM quality



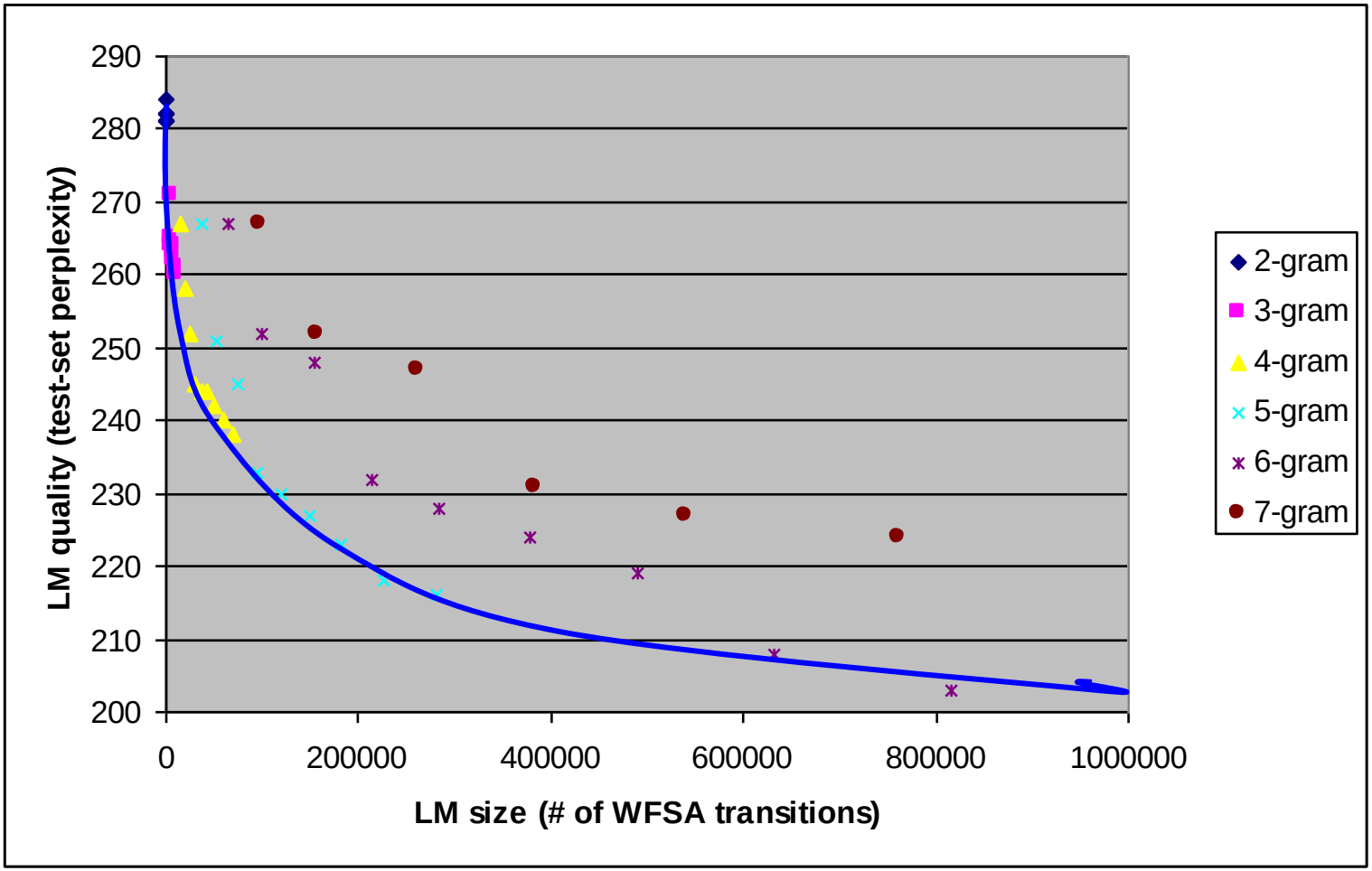
singletons pruned at all n-gram orders

```
% carmel -S test fsa
% carmel -c fsa
```

LM size vs. LM perplexity



LM size vs. LM perplexity



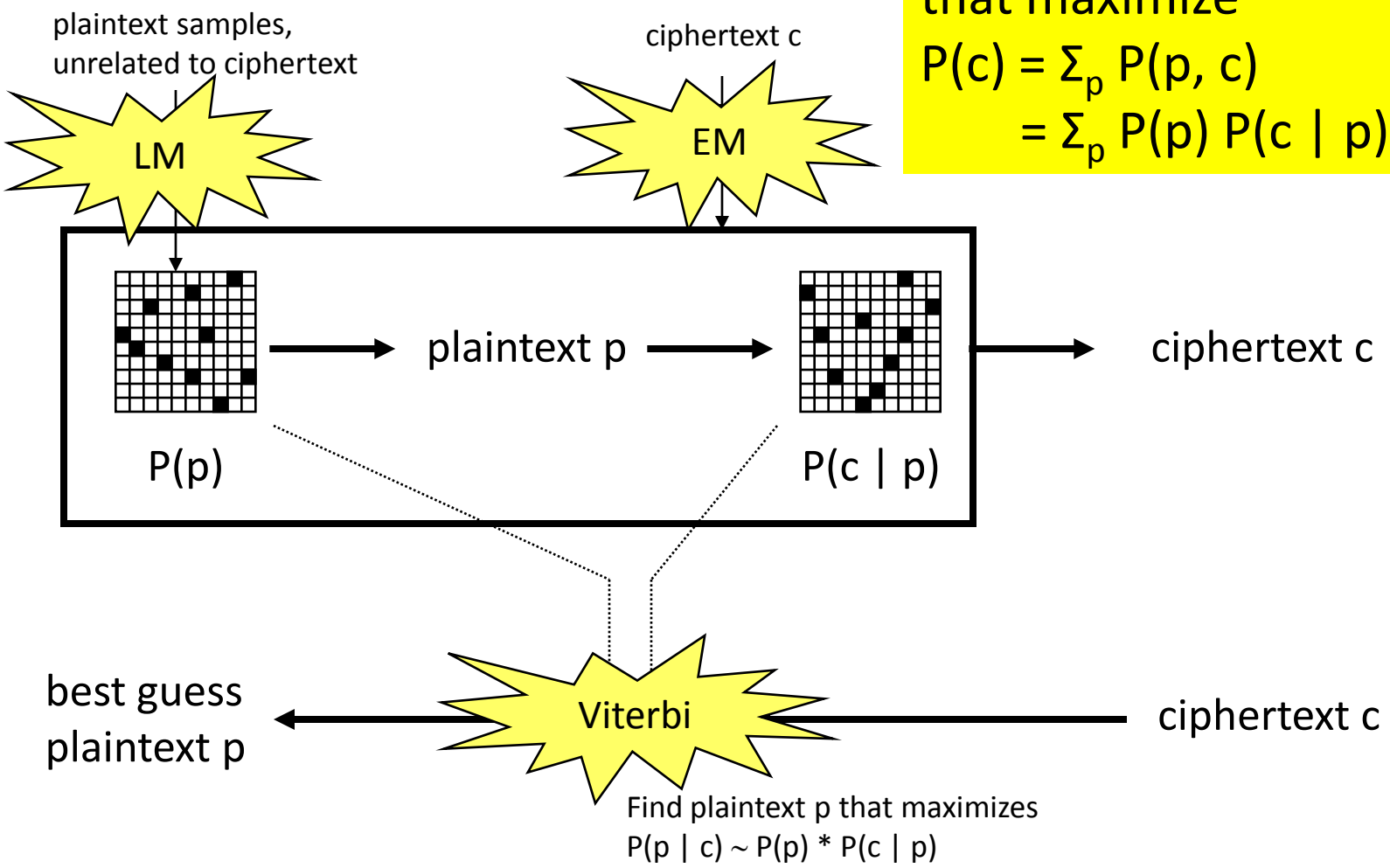
OK, we've got our 5-grams and 7-grams… back to decipherment

basic technical approach

Find substitution-table values that maximize

$P(c) = \Sigma_p P(p, c)$

$= \Sigma_p P(p) P(c \mid p)$



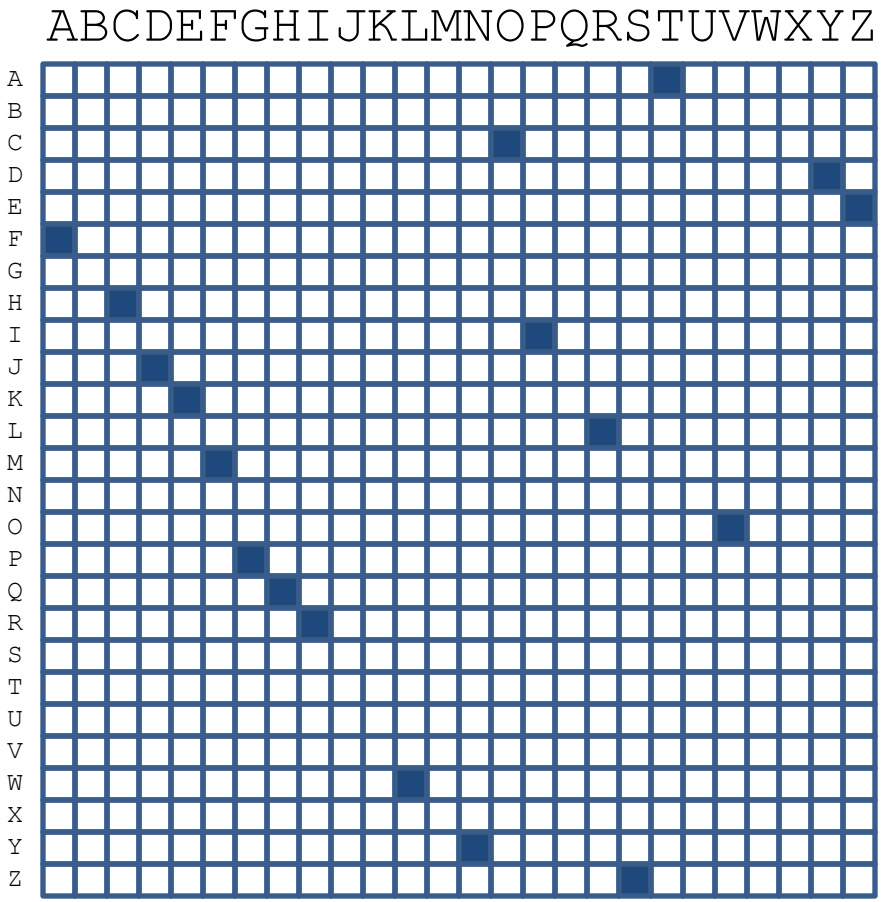
decipherment results: very bad

- 2-gram (41% error):
  - THE AVERAGE INCOULIZAN HAL BY WIED A MEVERENGE COR ANTULEXTY THAT HE POFOF MATHES BE PROND THAN MERENT WITIS PRASTHES
- 3-gram (60% error):
  - THE PREPACE EXCLUDICAL ING OF WEED A SUREMENCE WAS ANTIONITY THAT OF JABLY GOVIES OF THAND TION SEMENT DEVES CONSTING
- 5-gram (54% error):
  - THE COURAGE FORMATIONS HAD TO SEEK A REFERENCE AND COMMITTED THAT HE WOULD NATION OF BRING THAT DEVELY AFTER OCCUPIED
- 7-gram (66% error):
  - THE SPECIAL EXPERIENCE HAD ON JULY A REFERENCE AND COMMUNITY THAT HE HUMAN RIGHTS OF WOMEN THIS REPORT WOULD BRACKETS

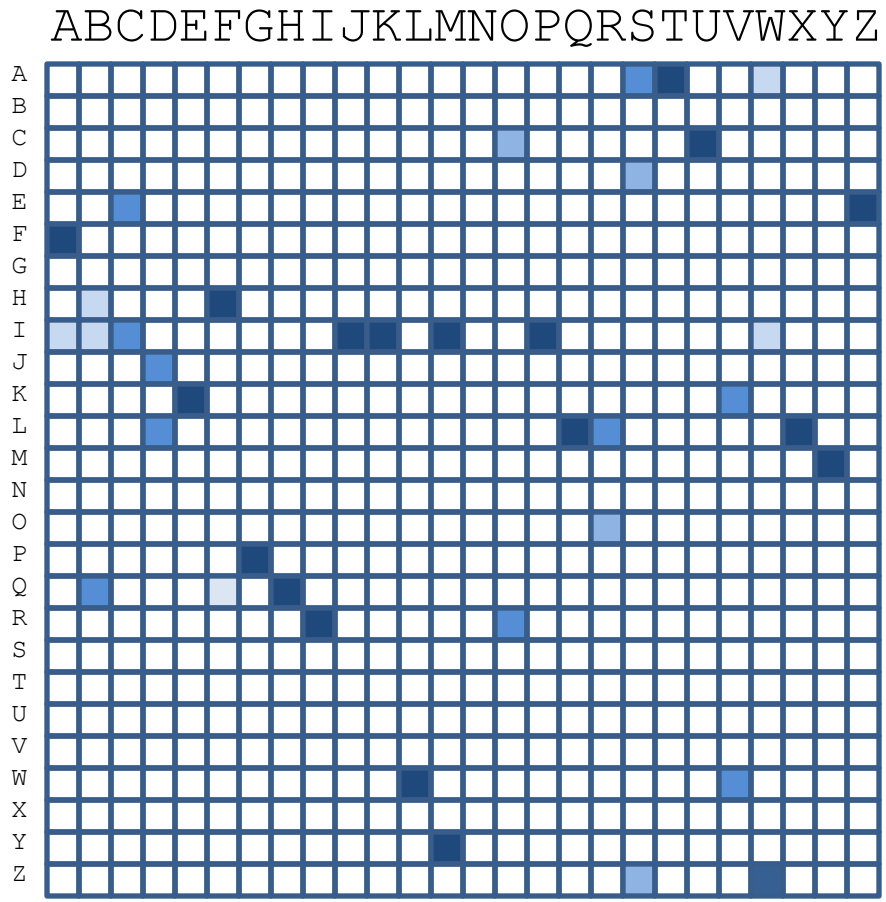
HOW CAN THIS BE HAPPENING?!

time to “dig in”

actual correct channel model



actual EM-learned channel model



time to “dig in”

Find substitution-table values that maximize

$$P(c) = \Sigma_p P(p, c)$$
$$= \Sigma_p P(p) \cdot P(c \mid p)$$

2-gram	P(best-plaintext)	P(cipher | best-plaintext)	P(cipher) = sum over all plaintexts of P(p) * P(c | p)	Decipher error
Correct answer	-282	0	-282	0
EM answer	-242	-33	-255	41%

```
THE AVERAGE INCOULIZAN HAL BY WIED A MEVERENGE COR ANTULEXTY THAT HE
POFOF MATHES BE PROND THAN MERENT WITIS PRASTHES

correct is:  THE AVERAGE ENGLISHMAN...
```

The 2-gram LM likes this string better than the correct decipherment (282 > 242).
EM is willing to pay the cost of a non-deterministic channel (33).
EM optimization criterion also sums over other plaintexts (242 + 33 > 255).

how can better LMs hurt?

2-gram	P(best-plaintext)	P(cipher | best-plaintext)	P(cipher) = sum over all plaintexts of P(p) * P(c | p)	Decipher error
Correct model	-282	0	-282	0
EM	-242	-33	-255	41%

5-gram	P(best-plaintext)	P(cipher | best-plaintext)	P(cipher) = sum over all plaintexts of P(p) * P(c | p)	Decipher error
Correct model	-241	0	-241	0
EM	-121	-84	-204	53%

```
THE COURAGE FORMATIONS HAD TO SEEK A REFERENCE AND COMMITTED
THAT HE WOULD NATION OF BRING THAT DEVELY AFTER OCCUPIED
```

5-gram model is even more opinionated that its decipherment is “good English”. It’s willing to pay for even more non-determinism in the channel.

change optimization criterion

- Even in training, we want the LM to "vote" less. Giving the channel a larger vote will encourage determinism.
  - Maximize: $P(e) \cdot P(c \mid e)^3$ when decoding
- Or: *cube-root* the LM probabilities before EM
  - had to settle for square root -- weak *awk* skills ☹

- Alternative:
  - encourage determinism via Bayesian methods
  - approximate Variational Bayes [Klein tutorial, ACL-07] did not work

change optimization criterion

5-gram	P(best-plaintext)	P(cipher | best-plaintext)	P(cipher) = sum over all plaintexts of P(p) * P(c | p)	Decipher error
Correct model	-241	0	-241	0
EM	-121	-84	-204	53%

```
THE COURAGE FORMATIONS HAD TO SEEK A REFERENCE AND COMMITTED
THAT HE WOULD NATION OF BRING THAT DEVELY - AFTER OCCUPIED
```

5-gram square root LM probs	P(best-plaintext)	P(cipher | best-plaintext)	P(cipher) = sum over all plaintexts of P(p) * P(c | p)	Decipher error
Correct model	-73	0	-73	0
EM	-42	-20	-60	11%

```
THE AVERAGE ENGLISHMEN HAS SO WEEK A REFERENCE FOR ANTIALITY
THAT HE WOULD RATHER BE PRONG THAN RECENT - DETER MCARTHUR
```

reduce LM vote during EM

N-gram order	Decipherment error	Decipherment error when reducing LM vote
2	41%	38%
3	59%	31%
5	53%	11%
7	65%	21%

Impact of the idea just discussed (square root LM probabilities)

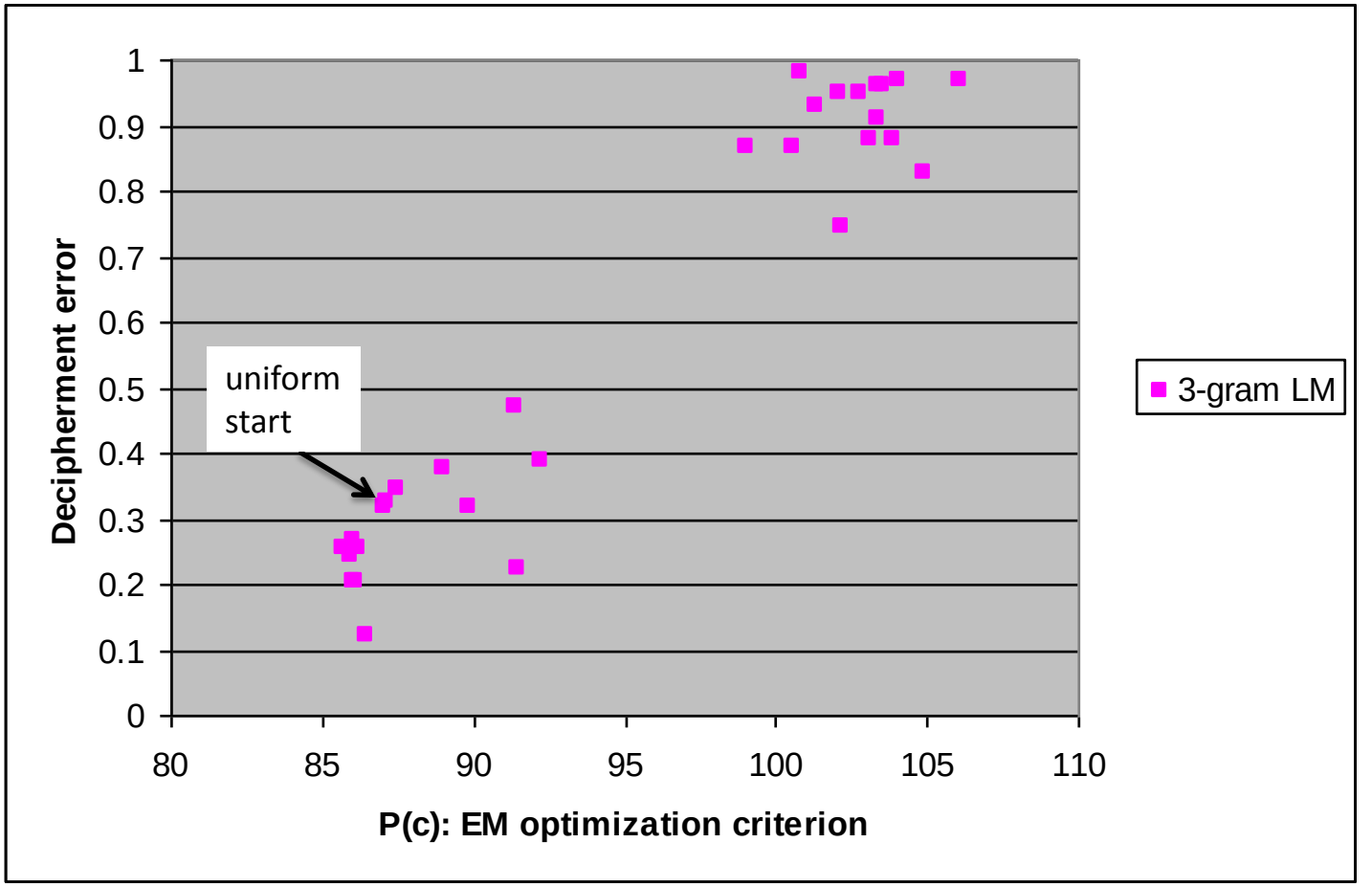
reduce LM vote during EM

N-gram order	Decipherment error	Decipherment error when reducing LM vote	With 10 random restarts
2	41%	38%	
3	59%	31%	
5	53%	11%	
7	65%	21%	11%

The new optimization criterion is working better, but EM is making a bad search error!

random restarts

98-letter cipher

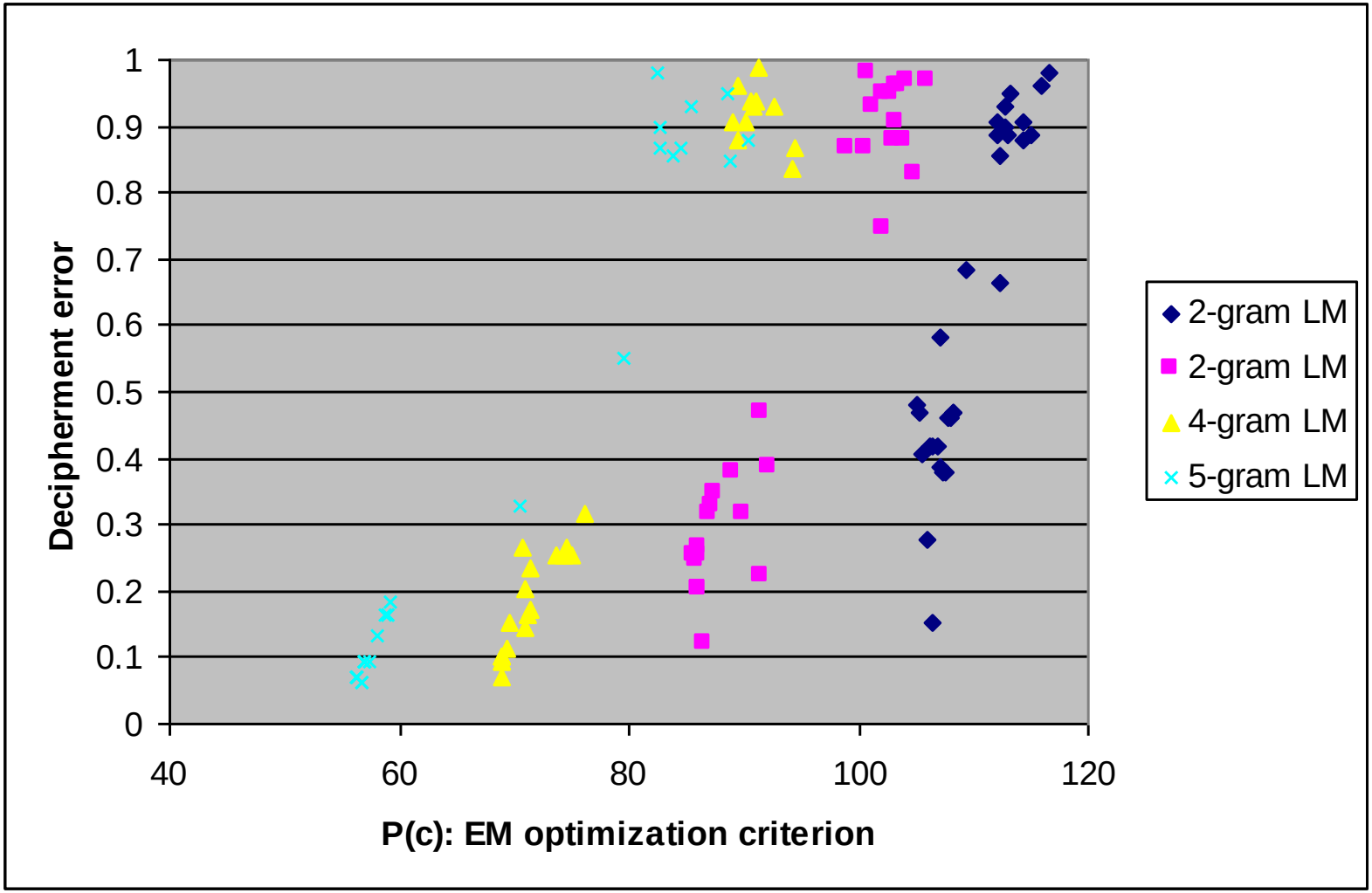


restart is like human using the pencil eraser to erase everything & start over

```
% carmel -t! 30 fsa2 cipher
```

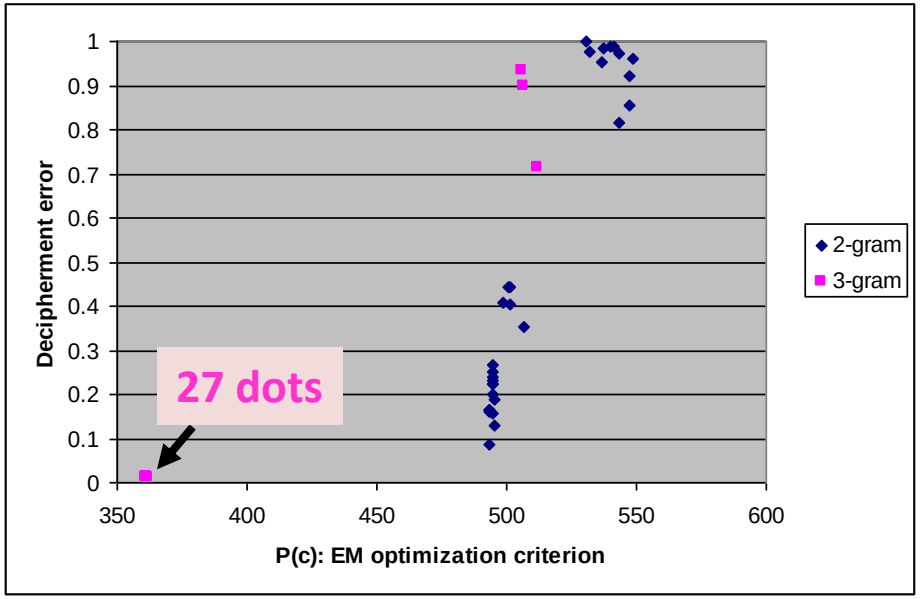
random restarts

98-letter cipher



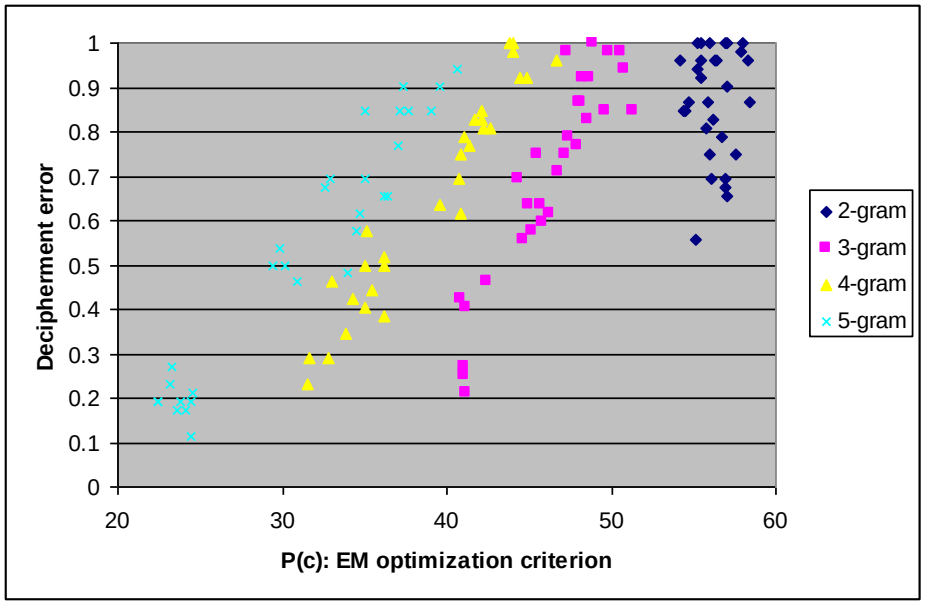
random restarts

EASY CIPHER (414 letters)



restarts not needed

HARD CIPHER (52 letters)



restarts badly needed

time for more LM data

N-gram order	Decipherment error	Decipherment error when reducing LM vote	With 10 random restarts	7m letters instead of 1.4m
2	41%	38%		
3	59%	31%		11%
5	53%	11%		6%
7	65%	21%	11%	7%

Brute force can now be applied

a niggling issue…

Manually set backoffs were:
{0.95, 0.90, 0.80, 0.70, 0.60, 0.20}
Empirically set backoffs are:
{0.22, 0.63, 0.62, 0.59, 0.54, 0.54}

N-gram order	Decipherment error	Decipherment error when reducing LM vote	With 10 random restarts	7m letters instead of 1.4m	Empirically estimated LM backoffs
2	41%	38%			
3	59%	31%		11%	
5	53%	11%		6%	
7	65%	21%	11%	7%	2%

```
czw qfwaqyw wiytobzhqi zqb
bn vwwg q awfwawiuw xna
qicoseocm czqc zw jnetv
aqczwa rw janiy czqi
awuwic gwcwa huqaczea
```

```
THE AVERAGE ENGLISHMAN HAS
SO DEEP A REVERENCE FOR
ANTIQUITY THAT HE WOULD
RATHER BE WRONG THAN
RECENT -PETER PLASTICS
```

```
THE AVERAGE ENGLISHMAN HAS
SO KEEP A REFERENCE FOR
ANTIQUITY THAT HE WOULD
RATHER BE WRONG THAN
RECENT -PETER MCARTHUR
```

Frequencies: “REFERENCE FOR” 496 “REVERENCE FOR” 1

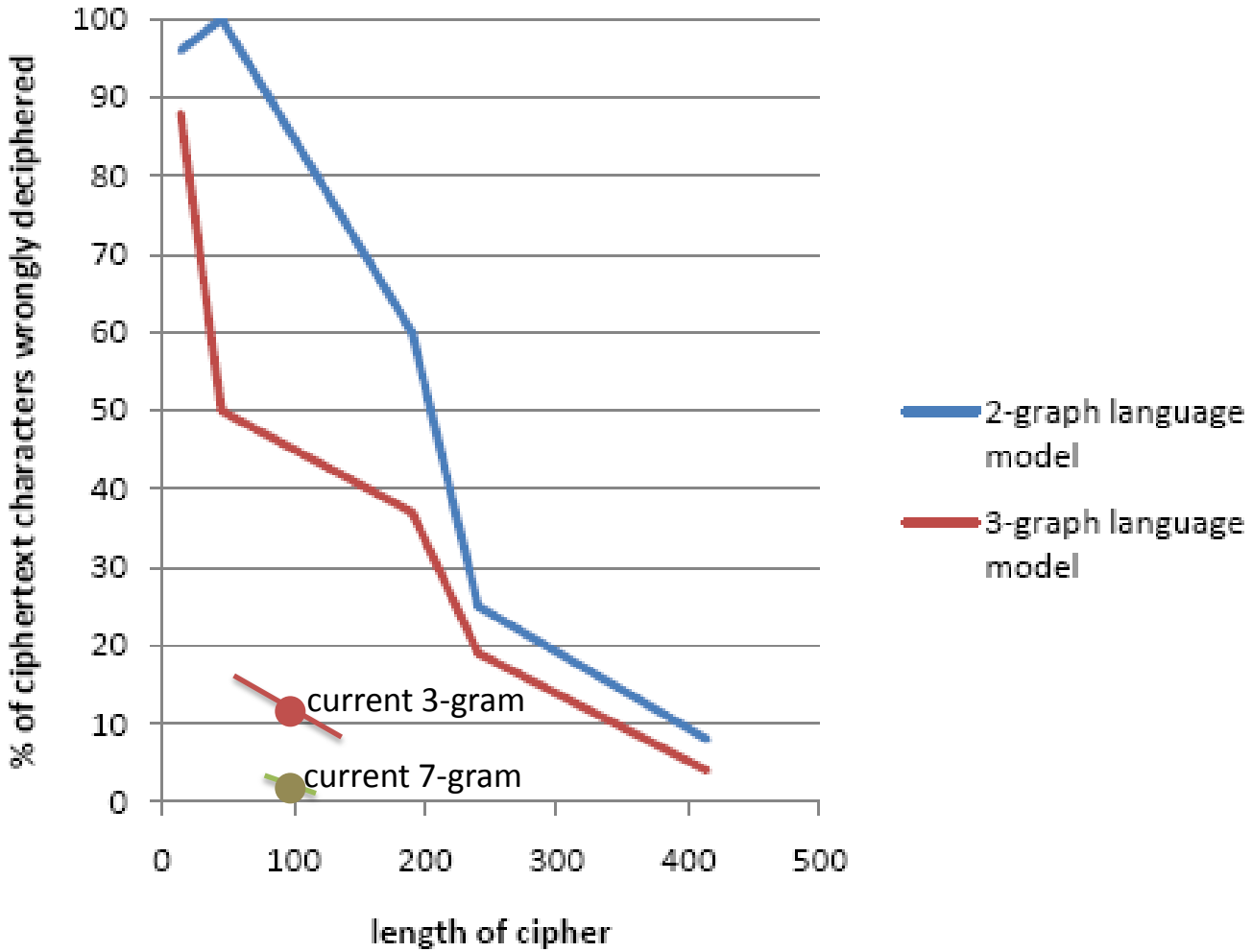
Zodiac killer cipher

(first of three parts)

Solution:

PT	CT
A	1 3 G l S
B	V
C	e
D	@ f
E	+ E I N P W Z
F	J Q
G	R
H	= M
I	Z k P U
J	
K	/
L	4 7 B
M	q
N	^ $ D O
O	; d T X
P	:
Q	
R	! \ r
S	1 3 6 F K
T	* H I L N
U	Y
V	]
W	A 2
X	t
Y	5
Z	

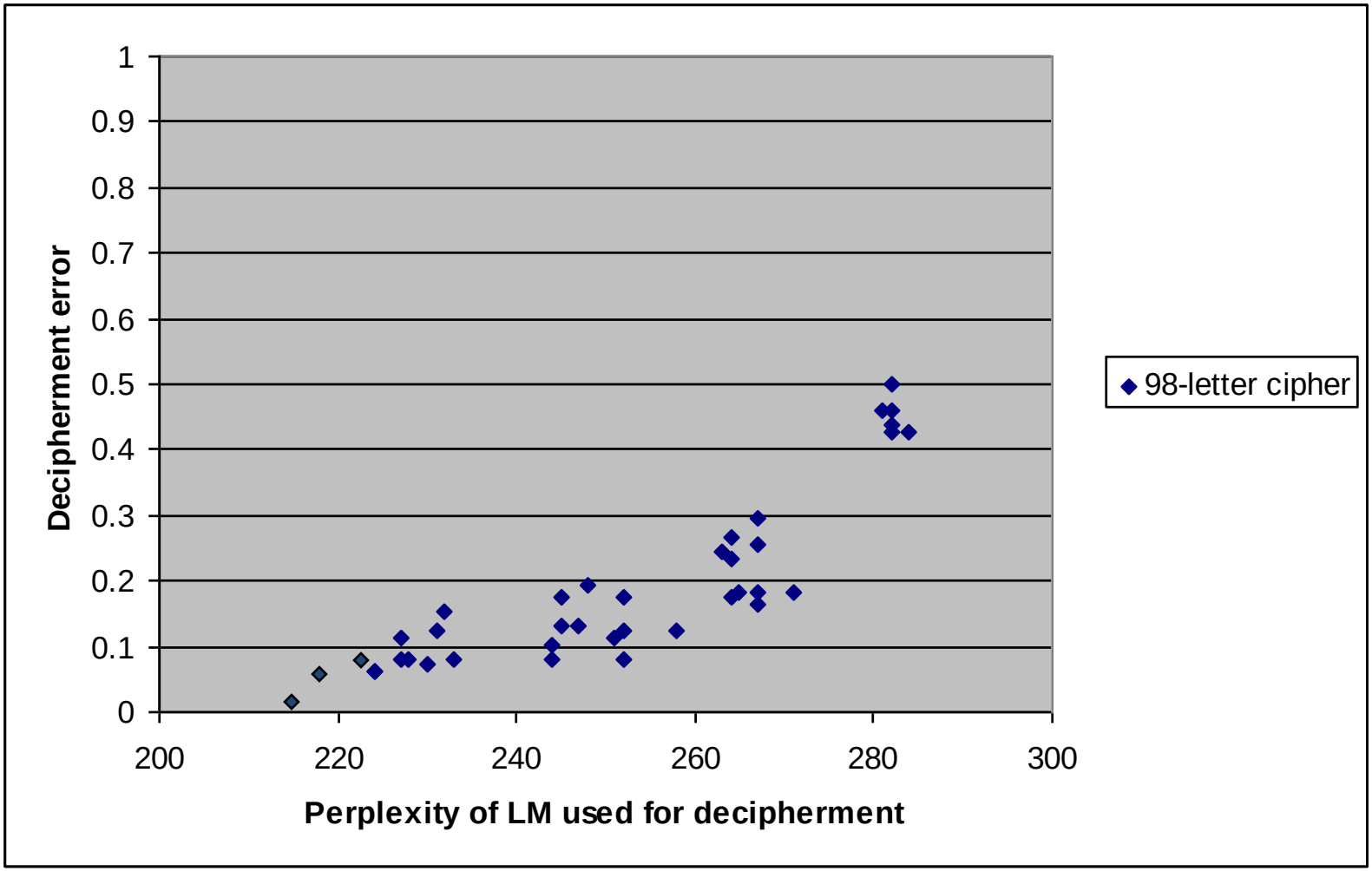
situation much more satisfactory



lessons

- make sure you can solve simple problems
- have a clear optimization criterion
- clarify search errors
- **when you get the wrong answer:**
  - **compare with right answer**
  - **why does wrong answer get a higher score?**
  - **break down the scoring components**
- use perplexity to pre-qualify knowledge resources

LM perplexity vs. task error



10 random restarts per point (best perplexity solution taken)

ok, are we done?

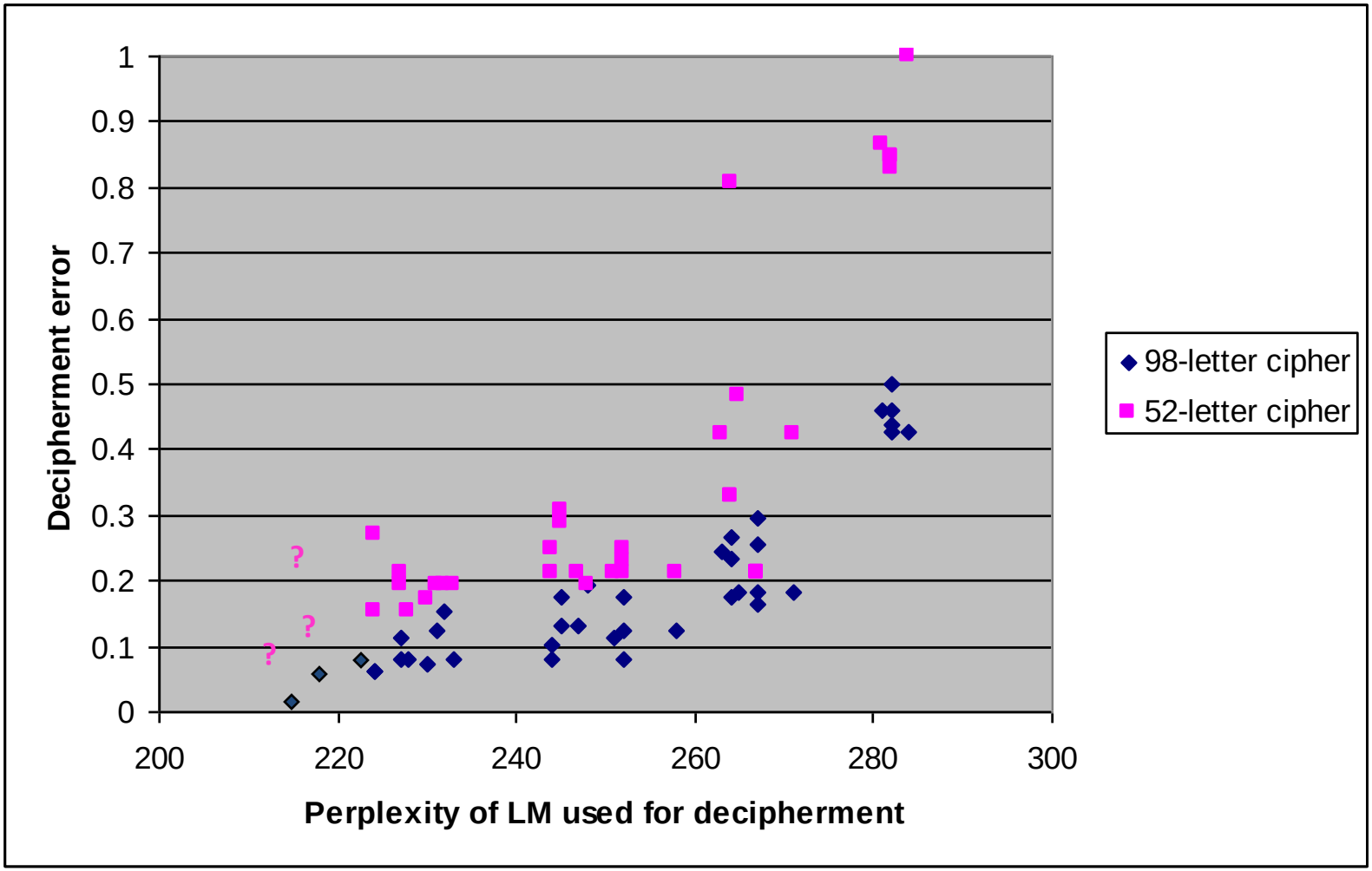
- Let’s try a 52-letter cipher, by truncating a longer one
- Ciphertext:
  - `pkxlygxf pjgfhp nlm gwwmrgnjm nfr mczogpgjm kxxvgfh xs bxofr`

ok, are we done?

- Let's try a 52-letter cipher, by truncating a longer one
- Ciphertext:
  - `pkxlygxf pjgfhp nlm gwwmrgnjm nfr mczogpgjm kxxvgfh xs bxofr`
- EM 7-gram (√), large training:
  - `SPORTION SHINGS ARE IMMEDIATE AND EXCLUSIVE WORKING OF BOUND`
- Correct answer:
  - `SCORPION STINGS ARE IMMEDIATE AND EXQUISITE COOLING OF WOUND`

EM solution has 19% error

LM perplexity vs. task error



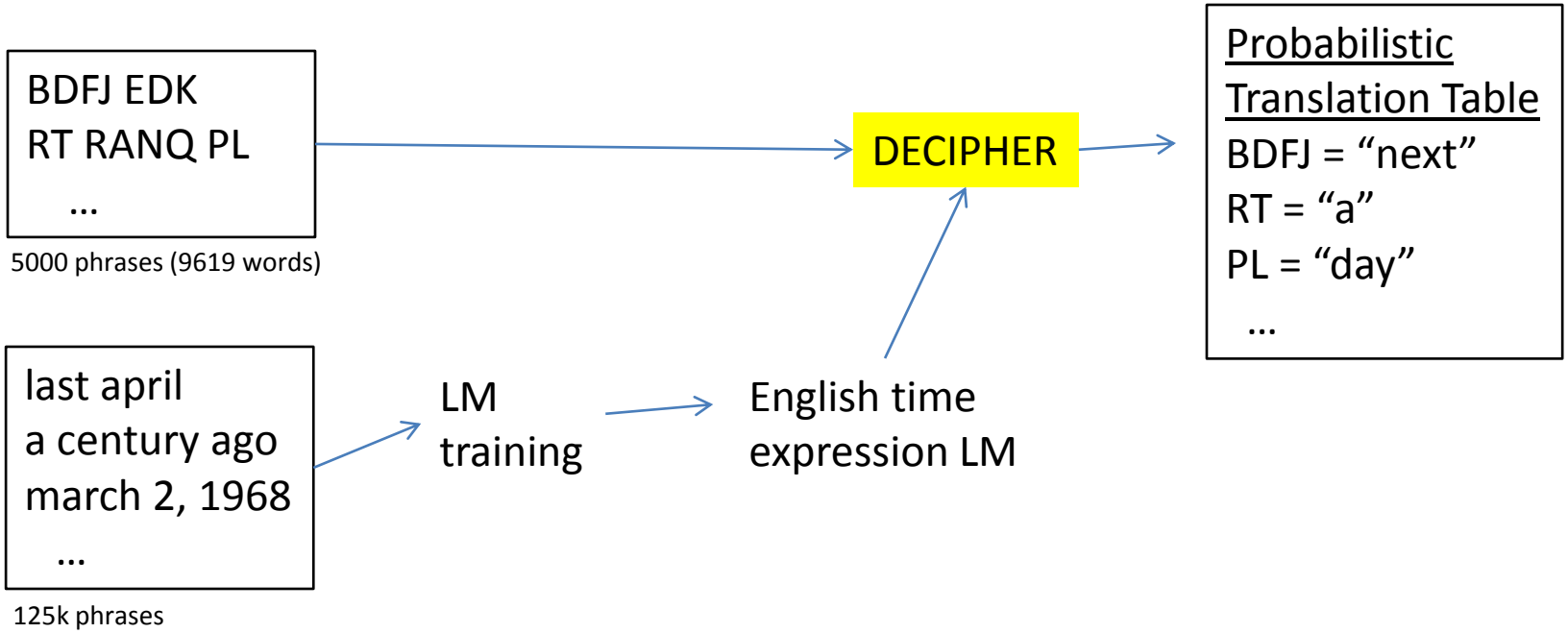
10 random restarts per point (best perplexity solution taken)

a few more experiments to report

- Can very powerful LMs solve very short ciphers?
  - Does a word unigram model have better perplexity than a letter 7-gram model? Is it more compact?
  - Could the power of a word-trigram model be applied in practice?
  - How much would a 1-for-1 constraint contribute?
- Can the method solve:
  - Ciphers with no spaces?
  - Homophonic ciphers? (multiple substitutions for frequent plaintext letters)
  - Phonetic substitution ciphers (name translation)?
  - Word substitution ciphers?
  - "Foreign language as a code for English" ciphers?
- What lessons for EM as we use it for GIZA++, category splitting, etc?

word substitution cipher

- Russian as a "code for English"
  - word substitution/transposition cipher
  - need to work up to this…



word substitution cipher: results

APRIL	april	0.41
APRIL	march	0.27
APRIL	july	0.25
APRIL	june	0.06
==========		
BEFORE	before	0.74
BEFORE	after	0.26
==========		
CENTURIES	decades	0.88
CENTURIES	days	0.04
CENTURIES	years	0.04
CENTURIES	weeks	0.03
==========		
CENTURY	century	0.96
CENTURY	decade	0.03
==========		
CONSECUTIVE	consecutive	0.79
CONSECUTIVE	full	0.18

DECEMBER	october	0.29
DECEMBER	february	0.26
DECEMBER	november	0.09
DECEMBER	december	0.03
DECEMBER	september	0.02
DECEMBER	friday	0.01
==========		
FEW	few	0.93
FEW	couple	0.04
FEW	dozen	0.03
==========		
FIVE	five	0.71
FIVE	nn	0.21
FIVE	eight	0.07
FIVE	hundred	0.01
==========		
JUST	only	0.59
JUST	just	0.41
==========		
WEEKENDS	november	0.54
WEEKENDS	days	0.37
WEEKENDS	october	0.07

After deciphering the 5000-word ciphertext:

943/9619 words wrong (9.8% error)

1-for-1 constraint

- back to letter substitution ciphers
- how can we add the **1-for-1 constraint**?
  - deterministic in the deciphering direction
  - deterministic in the enciphering direction
    - contrast with homophonic Zodiac cipher & NLP problems
- hard to convince EM to examine *only those solutions* that fit the constraint

integer programming

Maximize:

$2x + y$

Subject To:

$y + x < 6.9$

$y - x < 2.5$

$y > 1.1$

IP solver (lp_solve or CPLEX)

maximizing…

value of objective function = 10

x = 4

y = 2

integer programming

Maximize:

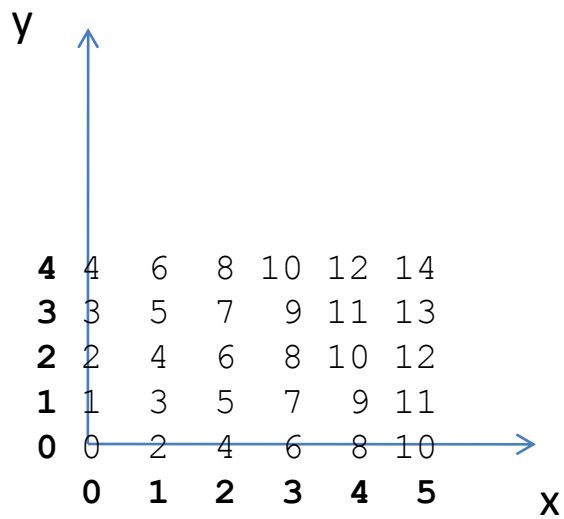
$2x + y$

Subject To:

$y + x < 6.9$

$y - x < 2.5$

$y > 1.1$



integer programming

Maximize:

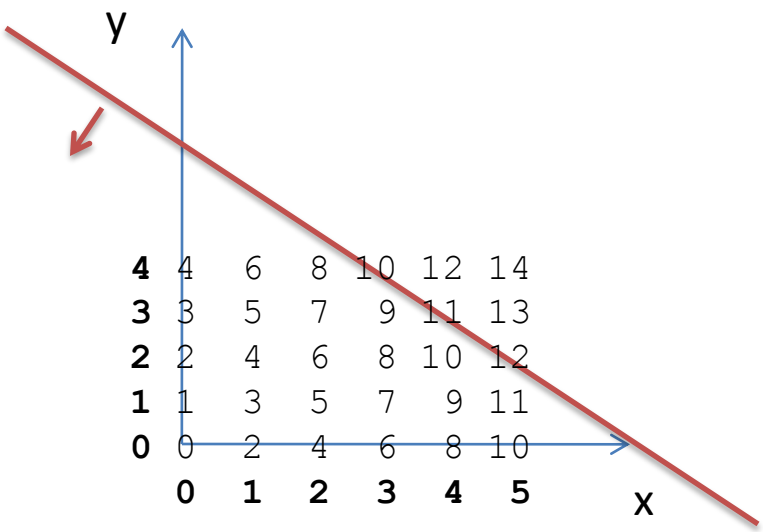
$2x + y$

Subject To:

$y + x < 6.9$

$y - x < 2.5$

$y > 1.1$



integer programming

Maximize:

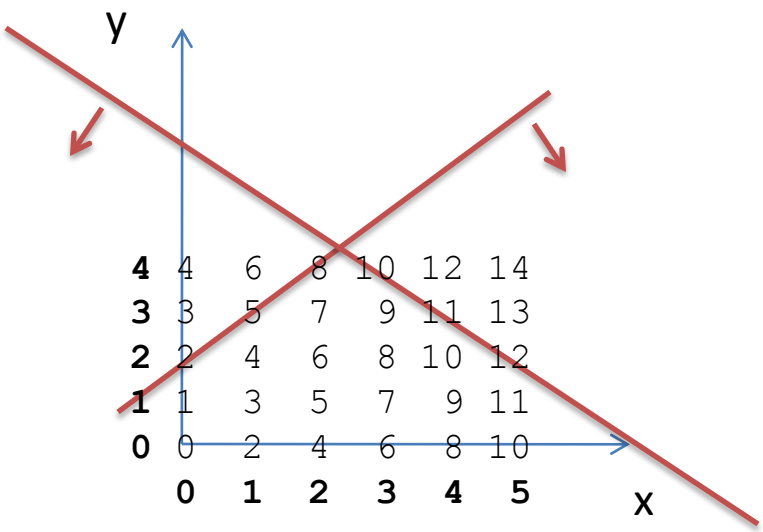
$2x + y$

Subject To:

$y + x < 6.9$

$y - x < 2.5$

$y > 1.1$



integer programming

Maximize:

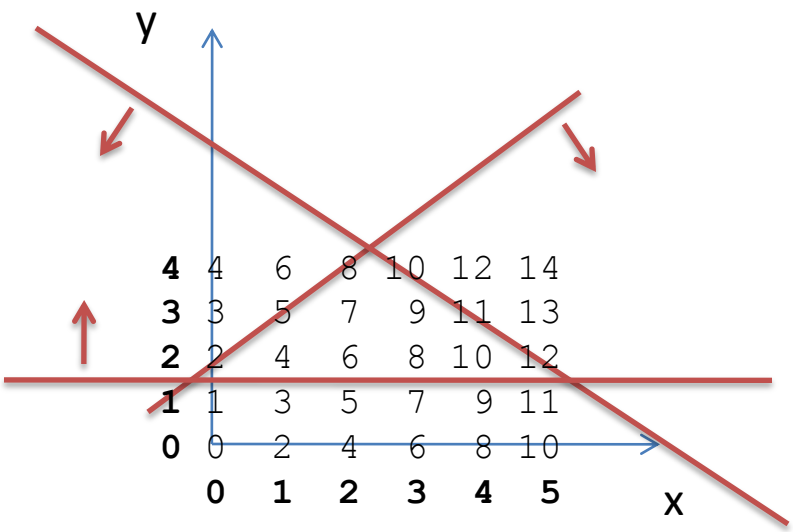
$2x + y$

Subject To:

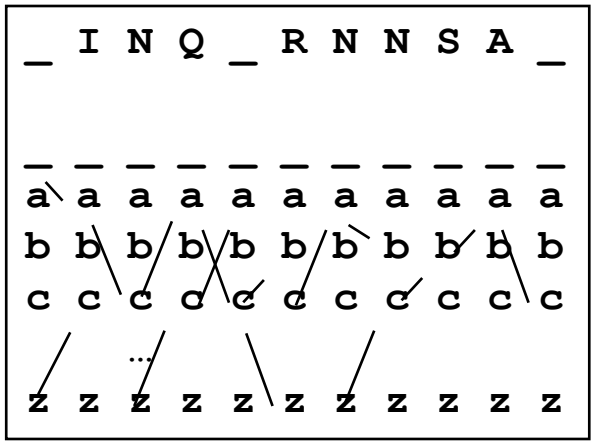
$y + x < 6.9$

$y - x < 2.5$

$y > 1.1$

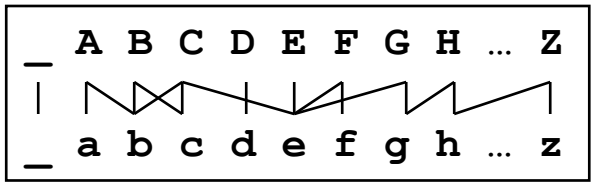


integer programming for letter substitution ciphers



every conceivable **link** is an "x" variable in the model (n x 27 x 27)

each link has a cost = -log P(y|x)



an additional 27x27 "y" variables make up the encipherment table

```
minimize:

0.068 x1_QU
+ 0.175 x1_Y_
+ 0.528 x1_VE
+ 0.572 x1_HE
+ 0.607 x1_D_
+ 0.717 x1_ZA
        ...
+ 0.068 x2_QU
+ 0.175 x2_Y_
+ 0.528 x2_VE
+ 0.572 x2_HE
+ 0.607 x2_D_
+ 0.717 x2_ZA
        ...
```

if this 2nd ciphertext letter is decoded as H and 3rd letter as E, then add cost –log P(E | H).

plaintext letter A must map to only one ciphertext letter

```
subject to:

r1: yAA + yAB + yAC + yAD + yAE + yAF
+ yAG + yAH + yAI + yAJ + yAK + yAL
+ yAM + yAN + yAO + yAP + yAQ + yAR
+ yAS + yAT + yAU + yAV + yAW + yAX
+ yAY + yAZ + yA_ = 1;
   ...
```

transform any ciphertext into integer program

```
#==========================================================
# Integer programming solution for 2-gram decipherment
# respecting 1-to-1 encipherment constraint
#
# Variables:
# - ypc:   1 iff <p,c> is an entry in the encipherment table
# - xi_ab: 1 iff ith letter is deciphered as a, and (i+1)st
#          letter is deciphered as b
#==========================================================

# minimize letter-bigram cost of solution

cp bb rrr.lp
echo ';' >>rrr.lp

awk '{for (j=1; j<=26; j++) let[j]=64+j;
      let[27] = 95; a=1;

# each plaintext letter maps to only one ciphertext letter

      for (j=1; j<=27; j++) {
        printf("r%d: ", a++);
        for (k=1; k<=27; k++)
          printf("+ y%c%c\n", let[j], let[k]);
        printf("= 1;\n")}

# each ciphertext letter maps to only one plaintext letter

      for (j=1; j<=27; j++) {
        printf("r%d: ", a++);
        for (k=1; k<=27; k++)
          printf("+ y%c%c\n", let[k], let[j]);
        printf("= 1;\n")}

# space maps to space

      printf("r%d: y__ = 1;\n", a++);

# path consistency.  links going in = links going out

      for (i=2; i<=NF-1; i++)
        for (j=1; j<=27; j++) {
          printf("r%d: ",a++);
          for (k=1; k<=27; k++)
            printf("+ x%d_%c%c\n", i-1, let[k], let[j]);
          printf("=\n");
          for (k=1; k<=27; k++)
            printf("+ x%d_%c%c\n", i, let[j], let[k]);
          printf(";\n")}

# first letter decodes as space

      printf("r%d: ",a++);
      for (k=1; k<=27; k++)
        printf("+ x1_%c%c\n", let[27], let[k]);
      printf("= 1;\n")

# last letter decodes as space

      printf("r%d: ",a++);
      for (k=1; k<=27; k++)
        printf("+ x%d_%c%c\n", NF-1, let[k], let[27]);
      printf("= 1;\n")

# each link sanctioned by encipherment table

      for (i=1; i<=NF-2; i++) {
        q=i+1;
        for (j=1; j<=27; j++) {
          printf("r%d: ",a++);
          for (k=1; k<=27; k++)
            printf("+ x%d_%c%c\n", i, let[k], let[j]);
          printf("= y%c%s;\n", let[j], $q)}}

# all variables integer

      printf("int ");
      for (i=1; i<=NF-1; i++)
        for (j=1; j<=27; j++)
          for (k=1; k<=27; k++)
            printf("x%d_%c%c,\n", i, let[j], let[k]);
      for (j=1; j<=27; j++)
        for (k=1; k<=27; k++)
          printf("y%c%c,\n", let[j], let[k]);
      printf("y__;\n")}' >>rrr.lp
```

solve integer program

(freeware “lp_solve”)

```
Variable y__ declared integer more than once, ignored on line 439700

Model name:  '' - run #1
Objective:   Maximize(R0)

SUBMITTED
Model size:     3297 constraints,   45198 variables,       134353 non-zeros.
Sets:                                   0 GUB,                 0 SOS.

Using DUAL simplex for phase 1 and PRIMAL simplex for phase 2.
The primal and dual simplex pricing strategy set to 'Devex'.

bfp_finishupdate: Failed at iter 98834, pivot 236;

LUSOL_RANKLOSS: Lost rank
   bfp_finishupdate: Correction or recovery was successful.
   coldual: Inaccurate bound-flip accuracy at iter 176460
   coldual: Inaccurate bound-flip accuracy at iter 176463
   coldual: Inaccurate bound-flip accuracy at iter 233026
                 ...
```

variables assigned value “1”

x1/_S	x21/_I	x41/SI
x2/SP	x22/IM	x42/IT
x3/PO	x23/MM	x43/TE
x4/OB	x24/ME	x44/E_
x5/BL	x25/ED	x45/_P
x6/LI	x26/DI	x46/PO
x7/IO	x27/IA	x47/OO
x8/ON	x28/AT	x48/OV
x9/N_	x29/TE	x49/VI
x10/_S	x30/E_	x50/IN
x11/ST	x31/_A	x51/NG
x12/TI	x32/AN	x52/G_
x13/IN	x33/ND	x53/_O
x14/NG	x34/D_	x54/OF
x15/GS	x35/_E	x55/F_
x16/S_	x36/EQ	x56/_C
x17/_A	x37/QU	x57/CO
x18/AB	x38/UR	x58/OR
x19/BE	x39/RI	x59/RN
x20/E_	x40/IS	x60/ND
		x61/D_

yAQ
yBR
yCU
yDV
yEW
yFX
yGY
yHZ
yIO
yJD
yKK
yLT
yMH
yNI
yON
yPG
yQS
yRA
ySB
yTC
yUE
yVF
yWJ
yXL
yYM
yZP

plaintext letter E is enciphered as W

19th cipher letter is deciphered as “B”, 20th letter as “E”.

Result on 52-letter cipher:

SPOBLION STINGS ABE IMMEDIATE AND EQURISITE POOVING OF CORND

21% error: compare with 85% error by EM and letter-bigram LM

Result on 98-letter cipher:

THE AVERAGE ENGLICHBAN HAC CO DEES
A REVERENME FOR ANTIQUITY
THAT HE POULD RATHER WE PRONG
THAN REMENT SETER BMARTHUR

12% error: compare with 45% error by EM and letter-bigram LM

Result on 414-letter cipher:

0.5 % error: compare with 10% error by EM and letter-bigram LM

lp_solve versus CPLEX

- Two different programs that solve integer programming problems

- 52-letter cipher
  - lp_solve: 6 hours
  - CPLEX: 3 minutes

- What's the difference between lp_solve and CPLEX?
  -

lp_solve versus CPLEX

- Two different programs that solve integer programming problems

- 52-letter cipher
  - lp_solve: 6 hours
  - CPLEX: 3 minutes

- What's the difference between lp_solve and CPLEX?
  - $990

comparing EM and IP

- Integer Programming (IP)
  - flexible constraints
  - slow
    - even CPLEX, on large ciphers with bigger n-grams
  - finds optimal solution
  - approximate solutions via “anytime” behavior
  - low memory requirement (RAM)

- EM
  - fast, linear-time per iteration
  - approximate solutions
  - improved solutions available via restarts
  - decipherment lattices created automatically by WFSA/WFST composition

- Final application -- word alignment

Spanish/English bilingual corpus

1a. Garcia and associates 1b. Garcia y asociados	7a. the clients and the associates are enemies 7b. los clientes y los asociados son enemigos
2a. Carlos Garcia has three associates 2b. Carlos Garcia tiene tres asociados	8a. the company has three groups 8b. la empresa tiene tres grupos
3a. his associates are not strong 3b. sus asociados no son fuertes	9a. its groups are in Europe 9b. sus grupos estan en Europa
4a. Garcia has a company also 4b. Garcia tambien tiene una empresa	10a. the modern groups sell strong pharmaceuticals 10b. los grupos modernos venden medicinas fuertes
5a. its clients are angry 5b. sus clientes estan enfadados	11a. the groups do not sell zenzanine 11b. los grupos no venden zanzanina
6a. the associates are also angry 6b. los asociados tambien estan enfadados	12a. the small groups are not modern 12b. los grupos pequenos no son modernos

Spanish/English bilingual corpus

1a. Garcia and associates 1b. Garcia y asociados	7a. the clients and the associates are enemies 7b. los clientes y los asociados son enemigos
2a. Carlos Garcia has three associates 2b. Carlos Garcia tiene tres asociados	8a. the company has three groups 8b. la empresa tiene tres grupos
3a. his associates are not strong 3b. sus asociados no son fuertes	9a. its groups are in Europe 9b. sus grupos estan en Europa
4a. Garcia has a company also 4b. Garcia tambien tiene una empresa	10a. the modern groups sell strong pharmaceuticals 10b. los grupos modernos venden medicinas fuertes
5a. its clients are angry 5b. sus clientes estan enfadados	11a. the groups do not sell zenzanine 11b. los grupos no venden zanzanina
6a. the associates are also angry	12a. the small groups are not modern

Idea: Considering all legal word alignments of this corpus, which one results in the **minimal-sized bilingual dictionary**? (i.e., the sparsest t-table).

Contrast with: Search for probabilistic t-table that maximizes Σ_a P(f,a | e) = ...
Print $\max_a$ P(f,a | e)

Spanish/English bilingual corpus

1a. Garcia and associates 1b. Garcia y asociados	7a. the clients and the associates are enemies 7b. los clientes y los asociados son enemigos
2a. Carlos Garcia has three associates 2b. Carlos Garcia tiene tres asociados	8a. the company has three groups 8b. la empresa tiene tres grupos
3a. his associates are not strong 3b. sus asociados no son fuertes	9a. its groups are in Europe 9b. sus grupos estan en Europa
4a. Garcia has a company also 4b. Garcia tambien tiene una empresa	10a. the modern groups sell strong pharmaceuticals 10b. los grupos modernos venden medicinas fuertes
5a. its clients are angry 5b. sus clientes estan enfadados	11a. the groups do not sell zenzanine 11b. los grupos no venden zanzanina
6a. the associates are also angry	12a. the small groups are not modern grupos pequenos no son modernos

Idea: Considering all legal word alignments of this corpus, which one results in the **minimal-sized bilingual dictionary**? (i.e., the sparsest t-table).

Spanish/English bilingual corpus

1a. Garcia and associates 1b. Garcia y asociados	7a. the clients and the associate 7b. los clientes y los asociados
2a. Carlos Garcia has three associates 2b. Carlos Garcia tiene tres asociados	8a. the company has three grou 8b. la empresa tiene tres grupos
3a. his associates are not strong 3b. sus asociados no son fuertes	9a. its groups are in Europe 9b. sus grupos estan en Europa
4a. Garcia has a company also 4b. Garcia tambien tiene una empresa	10a. the modern groups sell stro 10b. los grupos modernos vend
5a. its clients are angry 5b. sus clientes estan enfadados	11a. the groups do not sell zenz 11b. los grupos no venden zanz
6a. the associates are also angry	12a. the small groups are not m grupos pequenos no so

Idea: Considering all legal word alignments of this corpus, which one results in the **minimal-sized bilingual dictionary**? (i.e., the sparsest t-table).

Total dictionary size = 39

also/empresa
also/estan
and/y
angry/enfadados
are/estan
are/no
are/son
are/tambien
associates/asociados
a/tiene
carlos/carlos
clients/clientes
company/empresa
company/una
do/no
enemies/enemigos
europe/europa
garcia/garcia
groups/grupos
groups/modernos
groups/pequenos
has/tambien
has/tiene
his/sus
in/en
its/sus
modern/grupos
modern/modernos
not/son
not/venden
pharmaceuticals/fuertes
sell/venden
sell/zanzanina
small/grupos
strong/fuertes
strong/medicinas
the/la
the/los
three/tres

integer program

x_1_3_2 = 1 means:
in sentence pair 1,
3rd spanish word is linked
to 2nd english word.

minimize:

```
  x_also_asociados
+ x_also_empresa
+ x_also_enfadados
+ x_also_estan
+ x_also_garcia
+ x_also_los
+ x_also_tambien
+ x_also_tiene
+ x_also_una
+ x_and_asociados
+ x_and_clientes
+ x_and_enemigos
```

x_also_tambien = 1 means:
we say “also/tambien” is
“in the dictionary”; else not.

subject to:

```
r1: x_1_1_1 + x_1_1_2 + x_1_1_3 = 1;
r2: x_1_2_1 + x_1_2_2 + x_1_2_3 = 1;
r3: x_1_3_1 + x_1_3_2 + x_1_3_3 = 1;
    ...
r1001: x_1_1_1 + x_1_2_1 + x_1_3_1 <= 1;
r1002: x_1_1_2 + x_1_2_2 + x_1_3_2 <= 1;
r1003: x_1_1_3 + x_1_2_3 + x_1_3_3 <= 1;
   ...
r10001: x_1_1_1 - x_garcia_garcia <= 0;
r10002: x_1_1_2 - x_and_garcia <= 0;
r10003: x_1_1_3 - x_associates_garcia <= 0;
r10004: x_1_2_1 - x_garcia_y <= 0;
r10005: x_1_2_2 - x_and_y <= 0;
r10006: x_1_2_3 - x_associates_y <= 0;
r10007: x_1_3_1 - x_garcia_asociados <= 0;
  ...
```

variables assigned value “1”

x_three_tres
x_the_los
x_the_la
x_strong_fuertes
x_small_pequenos
x_sell_venden
x_pharmaceuticals_medicinas
x_not_no
x_modern_modernos
x_its_sus
x_in_europa
x_his_sus
x_has_tiene
x_groups_grupos
x_garcia_garcia
x_europe_en
x_enemies_enemigos
x_do_zanzanina
x_company_empresa
x_clients_clientes
x_carlos_carlos
x_a_una
x_associates_asociados
x_are_son
x_are_estan
x_angry_enfadados
x_and_y
x_also_tambien

x_9_5_4
x_9_4_5
x_9_3_3
x_9_2_2
x_9_1_1
x_8_5_5
x_8_4_4
x_8_3_3
x_8_2_2
x_8_1_1
x_7_7_7
x_7_6_6
x_7_5_5
x_7_4_4
x_7_3_3
x_7_2_2
x_7_1_1
x_6_5_5
x_6_4_3
x_6_3_4
x_6_2_2
x_6_1_1
x_5_4_4
x_5_3_3
x_5_2_2
x_5_1_1

x_4_5_4
x_4_4_3
x_4_3_2
x_4_2_5
x_4_1_1
x_3_5_5
x_3_4_3
x_3_3_4
x_3_2_2
x_3_1_1
x_2_5_5
x_2_4_4
x_2_3_3
x_2_2_2
x_2_1_1
x_1_3_3
x_1_2_2
x_1_1_1
x_12_6_6
x_12_5_4
x_12_4_5
x_12_3_2
x_12_2_3
x_12_1_1
x_11_5_3
x_11_4_5

x_11_3_4
x_11_2_2
x_11_1_1
x_10_6_5
x_10_5_6
x_10_4_4
x_10_3_2
x_10_2_3
x_10_1_1

in sentence pair 12, there is a link between spanish position 4 and english position 5.

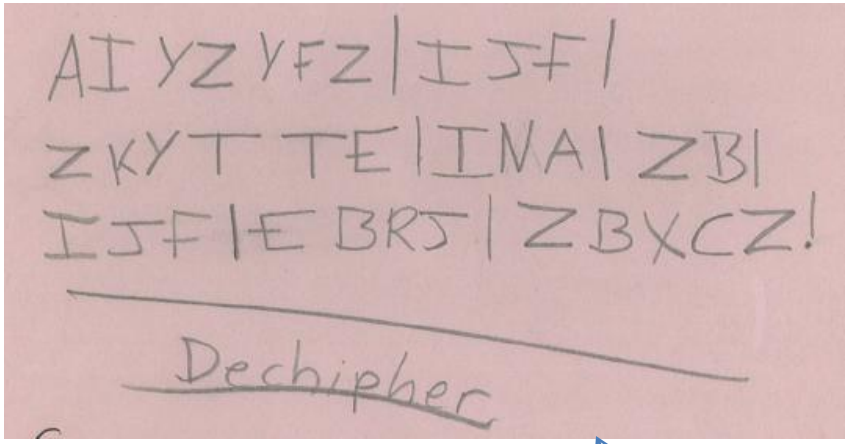
integer program solution

1a. Garcia and associates 1b. Garcia y asociados	7a. the clients and the associates are enemies 7b. los clientes y los asociados son enemigos
2a. Carlos Garcia has three associates 2b. Carlos Garcia tiene tres asociados	8a. the company has three groups 8b. la empresa tiene tres grupos
3a. his associates are not strong 3b. sus asociados no son fuertes	9a. its groups are in Europe 9b. sus grupos estan en Europa
4a. Garcia has a company also 4b. Garcia tambien tiene una empresa	10a. the modern groups sell strong pharmaceuticals 10b. los grupos modernos venden medicinas fuertes
5a. its clients are angry 5b. sus clientes estan enfadados	11a. the groups do not sell zenzanine 11b. los grupos no venden zanzanina
6a. the associates are also angry 6b. los asociados tambien estan enfadados	12a. the small groups are not modern 12b. los grupos pequenos no son modernos

summary

- several unsupervised decipherment tasks
- LM size vs. LM quality
- LM quality vs. task error
- modified optimization criterion for EM
- integer programming for global constraints
- word alignment application

thɛ : ɛnd



challenge from
Victoria Knight

thε : εnd

AI YZ YFZ | IJF |
ZKYT TE | INA | ZB |
IJF | EBRJ | ZBXCZ!

Dechipher

tt zz
ff ll mm pp

AIYZYFZ IJF ZKYTTE INA ZB
Daisies are $mpLLy and $o
IJF EBRJ ZBXCZ
are your $ocks$

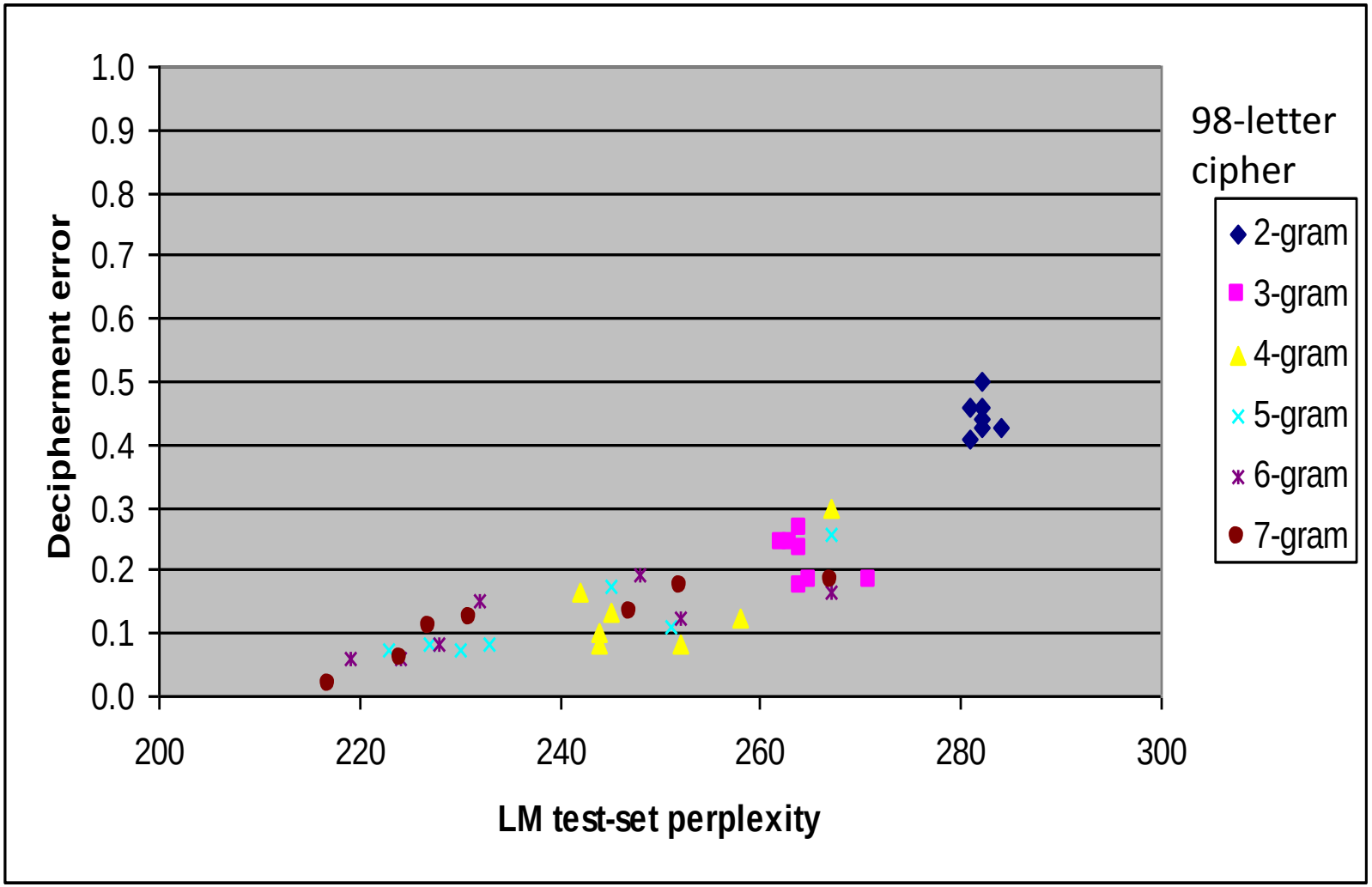
IJF the
INA the
ZB in of an to
E

A ||
B |||
C |
D
E ||
F ||||
G
H
I ||||
J |||
K |
L
M
N |
O
P
Q
R |
S
T ||
U
V
W
X |
Y ||||
Z ||||| |

a ✓
b
c
d ✓
e ✓
f
j
h
i ✓
j
k
l
m
n ✓
o ✓
p
q
r ✓
s ✓
t
u

_ _ntnet
_t _s_es
~~_tstes~~
_ _isies
s_itt_
s_iff_
s_imm_

LM perplexity vs. task error



10 random restarts per point (best perplexity solution taken)