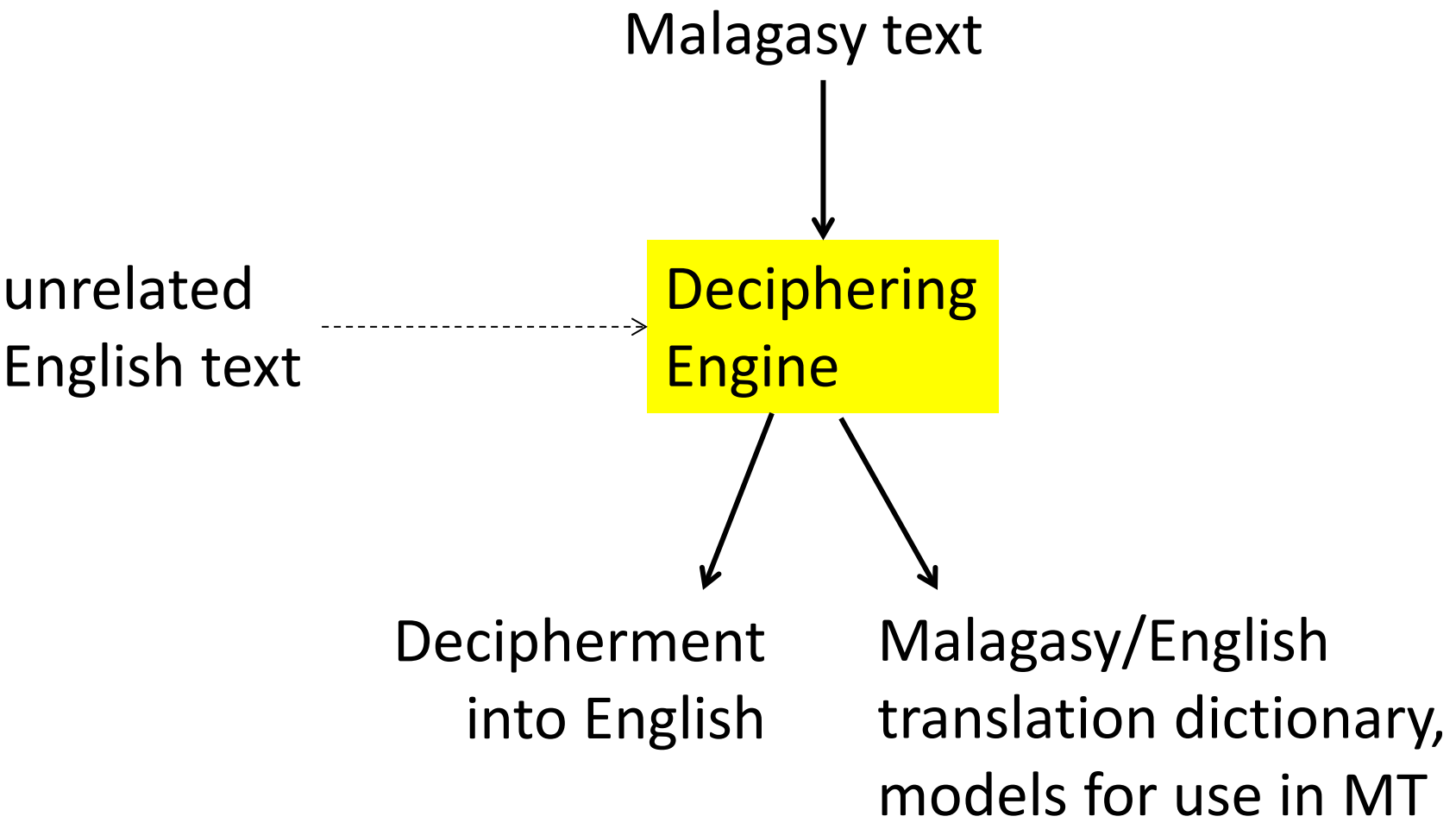
Deciphering Malagasy

cross-site collaboration (ISI/UT/CMU/MIT):
using **dependency parsers**
to extract **translation patterns**
from **non-parallel text**

Qing Dou & Kevin Knight
USC/ISI

MURI Review
November 22, 2013

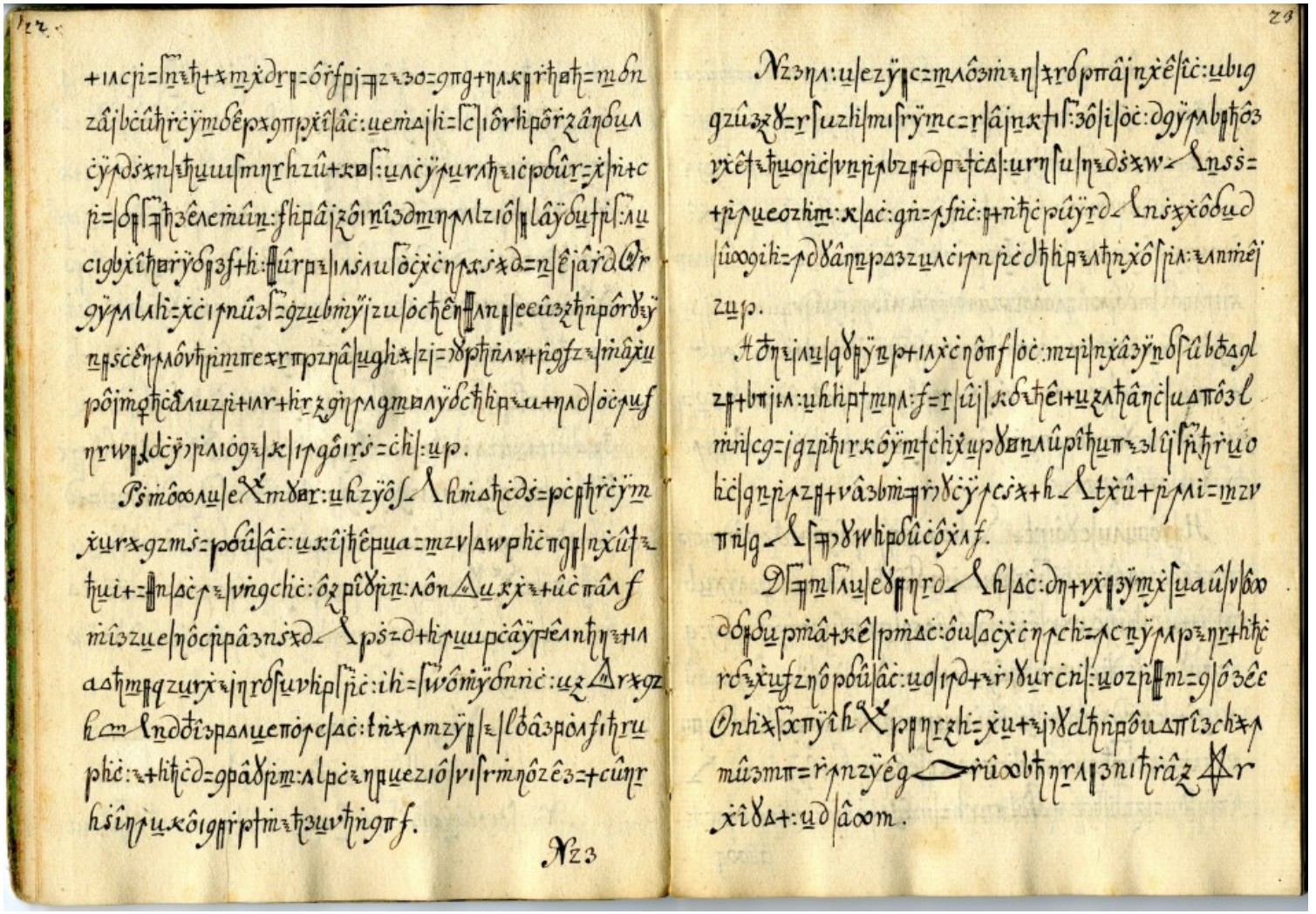
Exploit Non-Parallel Data for MT



Outline

- View foreign text as a code for English
- Decipher it
  - using any linguistic patterns we need
- Step-by-step
  - letter substitution cipher
  - word substitution cipher
  - foreign language as cipher

Letter Substitution Cipher



KDCY LQZKTLJKX CY MDBCYJQL: "TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

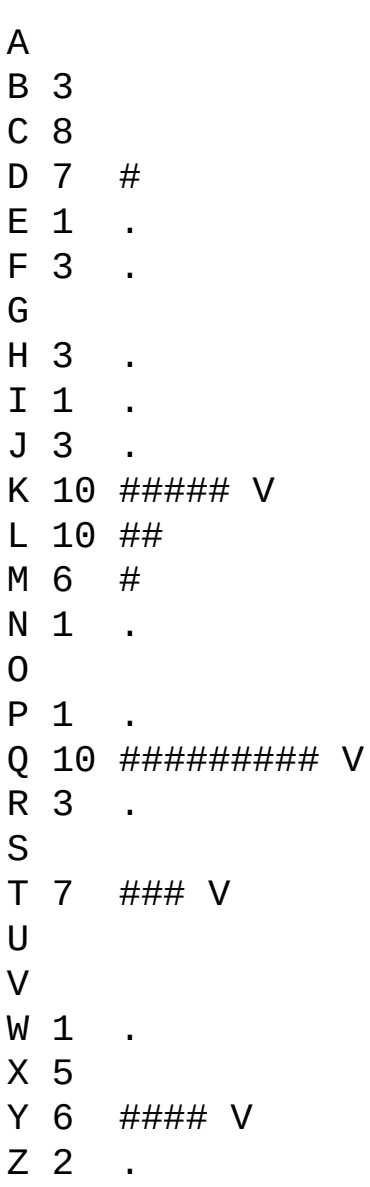
PTBMYEQL K FKH CY LQZKTL TC."

auto repairman to customer: if
KDCY LQZKTLJKX CY MDBCYJQL: "TR

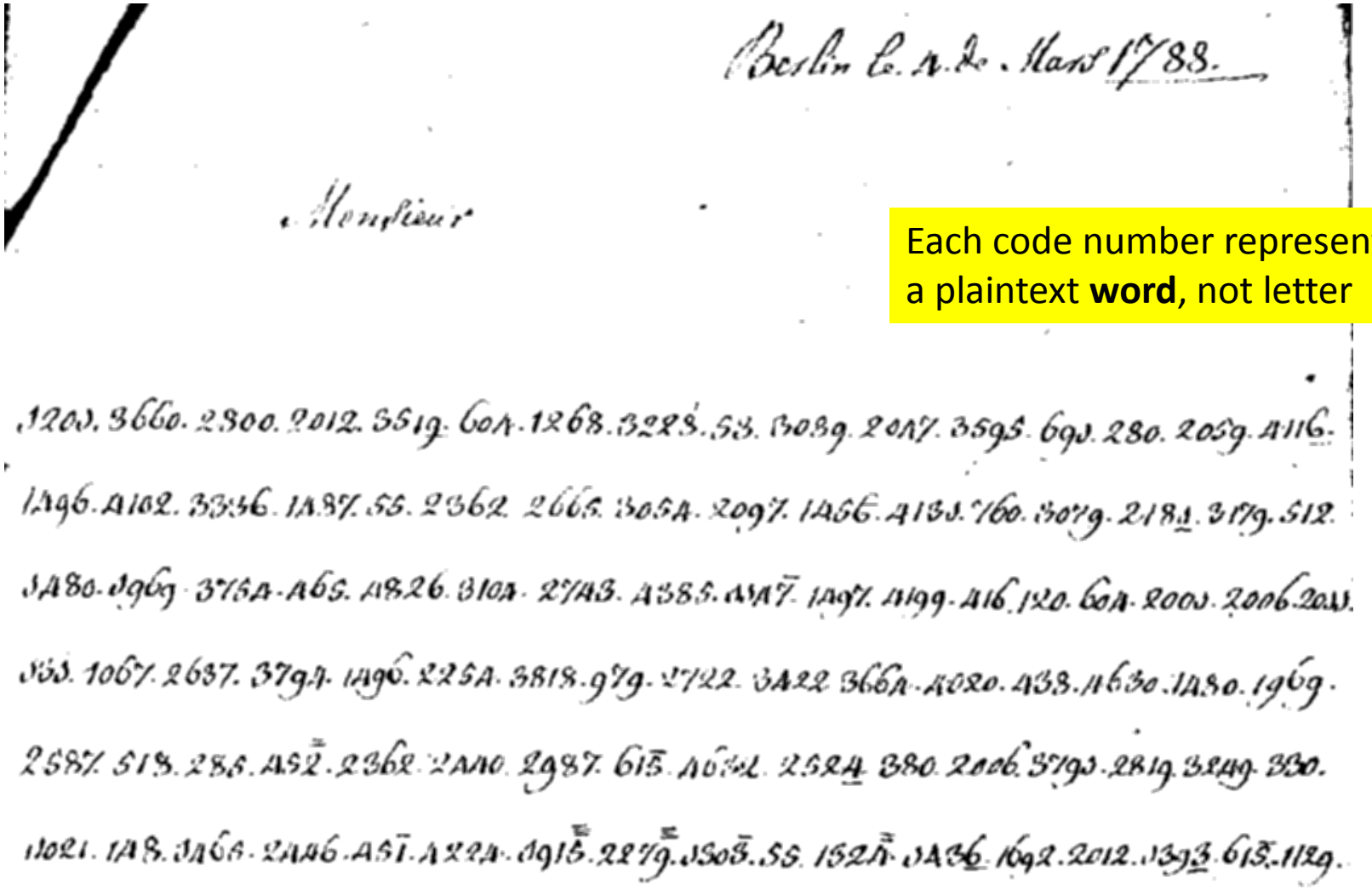
you wait we can freeze your
HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics
MKL DXCTW RDCDLQ JQMNKXTMB

discover a way to repair it
PTBMYEQL K FKH CY LQZKTL TC."



Word Substitution



Berlin le 2. Mars 1788.

Monsieur

Each code number represents a plaintext **word**, not letter

1203. 3660. 2300. 2012. 3519. 604. 1268. 3228. 53. 3089. 2017. 3595. 693. 280. 2059. 4116.

1496. 4102. 3336. 1487. 55. 2362. 2665. 3054. 2097. 1456. 4133. 760. 3079. 2181. 3179. 512.

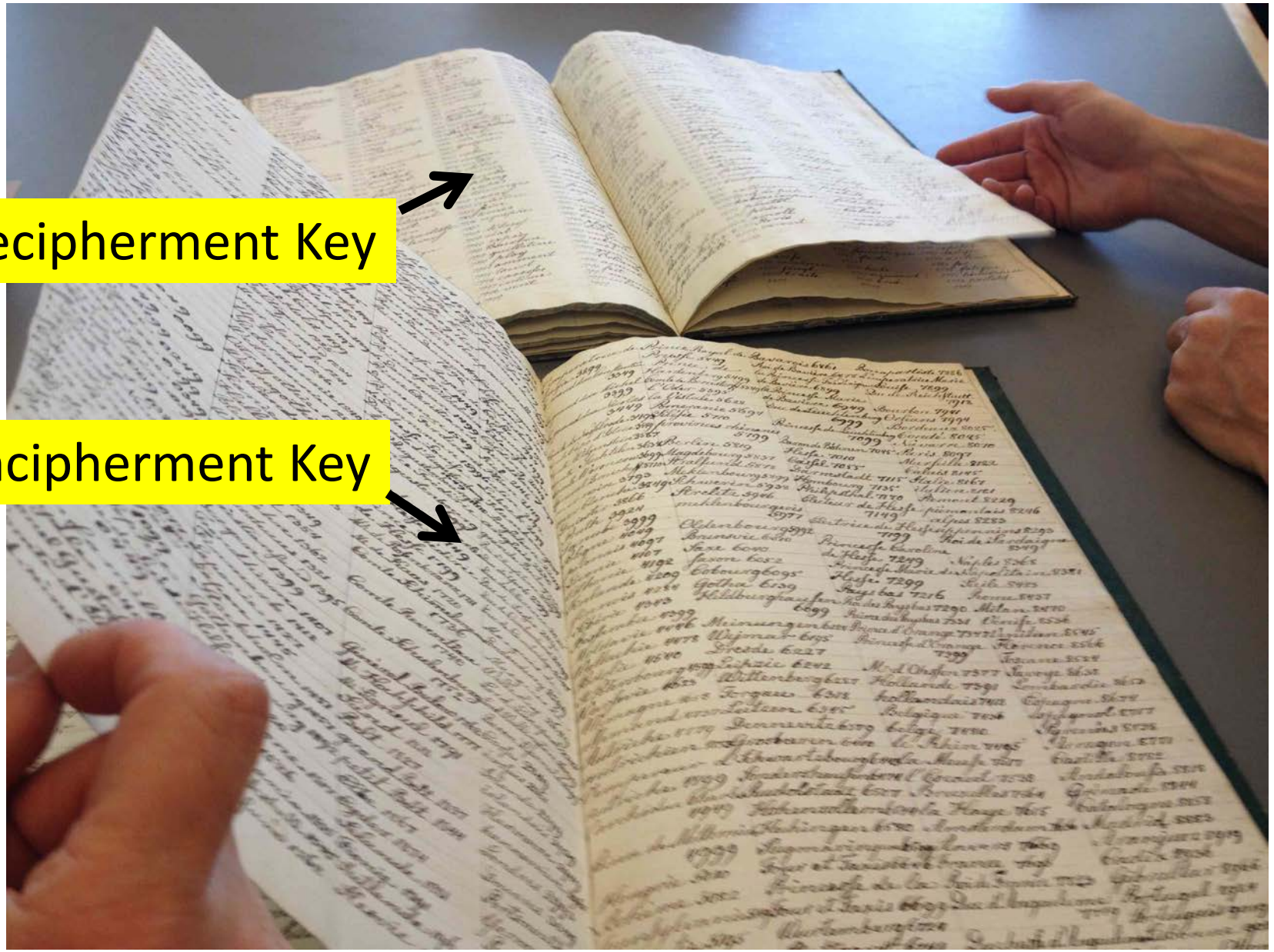
3480. 3969. 3754. 465. 4826. 3104. 2743. 4385. 4147. 1497. 4199. 416. 120. 604. 2003. 2006. 203.

333. 1067. 2637. 3794. 1496. 2254. 3818. 979. 2722. 3422. 3664. 4020. 433. 4630. 1480. 1969.

2587. 518. 286. 452. 2362. 2410. 2987. 615. 4642. 2524. 380. 2006. 3793. 2819. 3249. 330.

1021. 148. 3466. 2446. 451. 4224. 3915. 2279. 3503. 55. 1524. 3436. 1692. 2012. 393. 615. 1129.

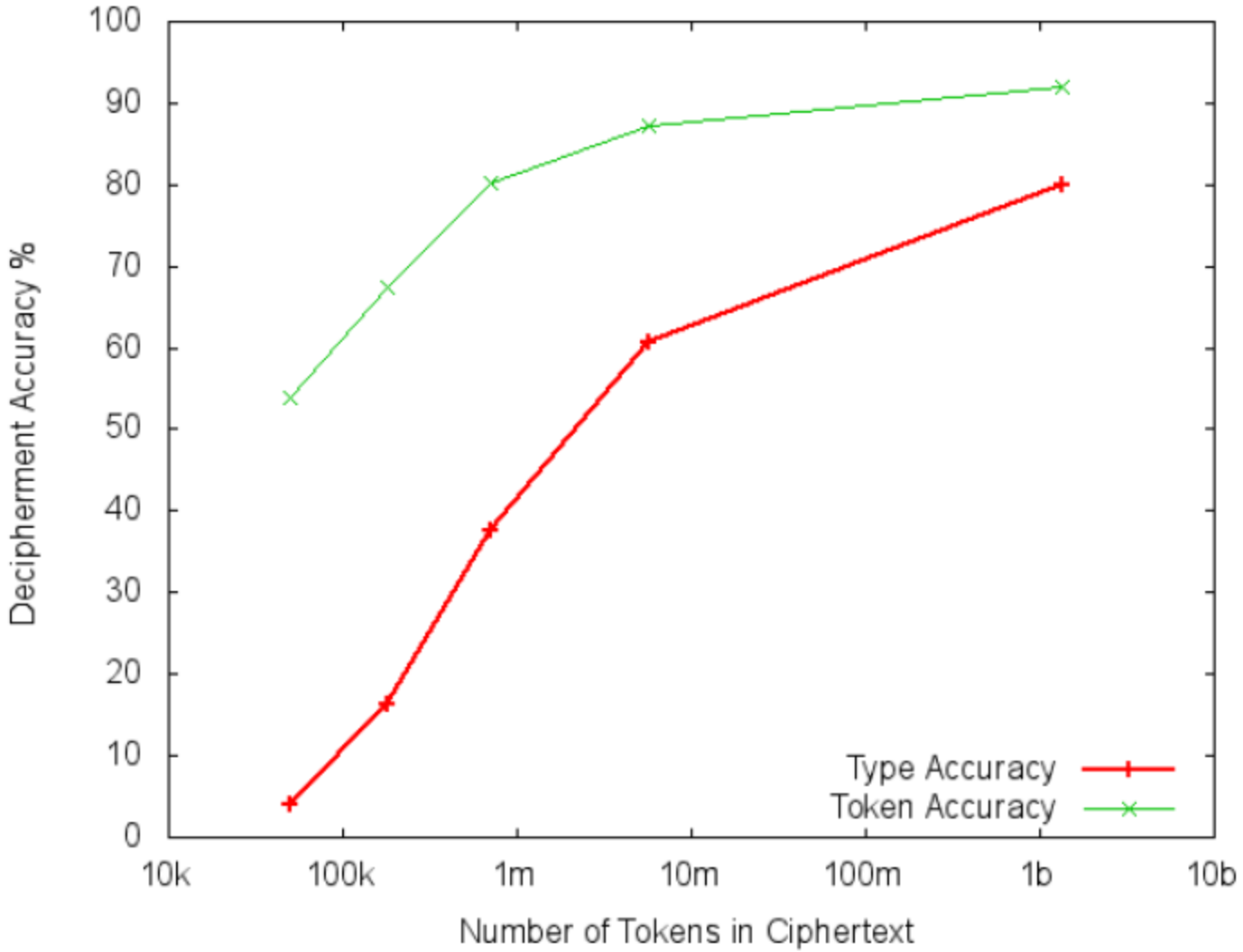
Word Substitution Keys



Word Substitution

- Suppose I replace each English word on your hard drive with some integer
- Can you recover your texts?
  - What key, if applied to your hard drive, would yield sensible English?
- In principle, we apply the same techniques we used for letter substitution

Word Substitution



(Dou & Knight 2012)

Foreign Language as a Code for English

!l@!m
!lywm
!lth!ny&
!l@!m !lm!Dy
Sfr
@!m
th!ny&
@!m 1992
@!m 1993
ywm
!l!sbw@ !lm!Dy
fy !ldqyq&
!lsn& !lj!ry&
!lsn&
!lsh=hr !lm!Dy
!lsh=hr !lj!ry
snw!t
sn&
=hdh! !l@!m
s!@&
!l@Sr
@!m 1991

@!m 1990
w!lth!ny&
fy !lywm
mn !lsh=hr !lj!ry
!lqrn
!'y!m
@!m!aN
!ls!@&
17 shb!T 1994
thl!th snw!t
dqyq&
=hdh=h !lsn&
ywmyn
mn !l@!m !lm!Dy
!lsn& !lmqbl&
fy !lsn&
kl ywm
fy !l@!m !lm!Dy

!l@Swr
=hdh! !lsh=hr
fy ywm
nys!n
!sbw@
=hdh=h !l!'y!m
qbl !'y!m
fy !l@Sr
mn !lsn&
!lsnw!t
b@d ywm
!l!y!m
13 nys!n 1994
!lth!ny& @sh!&
thl!th& !y!m
qbl !sbw@yn
fy !lywm !lt!ly
sh@b!n
tmwz
3 dhw !lHj& 1414
fy shb!T !lm!Dy
qbl ywmyn

Foreign Language as a Code for English

<n> Hzyr!n <n>

13	4 Hzyr!n 1967	2	fy 30 Hzyr!n 1995
12	fy 12 Hzyr!n 1993	2	fy 18 Hzyr!n 1994
7	5 Hzyr!n 1967	2	fy 14 Hzyr!n 1993
6	fy 30 Hzyr!n 1989	2	fy 14 Hzyr!n 1991
6	30 Hzyr!n 1989	2	fy 12 Hzyr!n 1990
4	fy 30 Hzyr!n 1994	2	7 Hzyr!n 1994
4	fy 30 Hzyr!n 1993	2	6 Hzyr!n 1941
3	fy 19 Hzyr!n 1967	2	26 Hzyr!n 1994
2	ywm 30 Hzyr!n 1989	2	21 Hzyr!n 1994
2	w 6 Hzyr!n 1994	2	1 Hzyr!n 1994
2	qbl 5 Hzyr!n 1967	2	19 Hzyr!n 1965
2	fy 9 Hzyr!n 1967	2	18 Hzyr!n 1994
2	fy 7 Hzyr!n 1981	2	18 Hzyr!n 1940
2	fy 6 Hzyr!n 1994	2	12 Hzyr!n 1993
2	fy 5 Hzyr!n 1967	2	11 Hzyr!n 1994

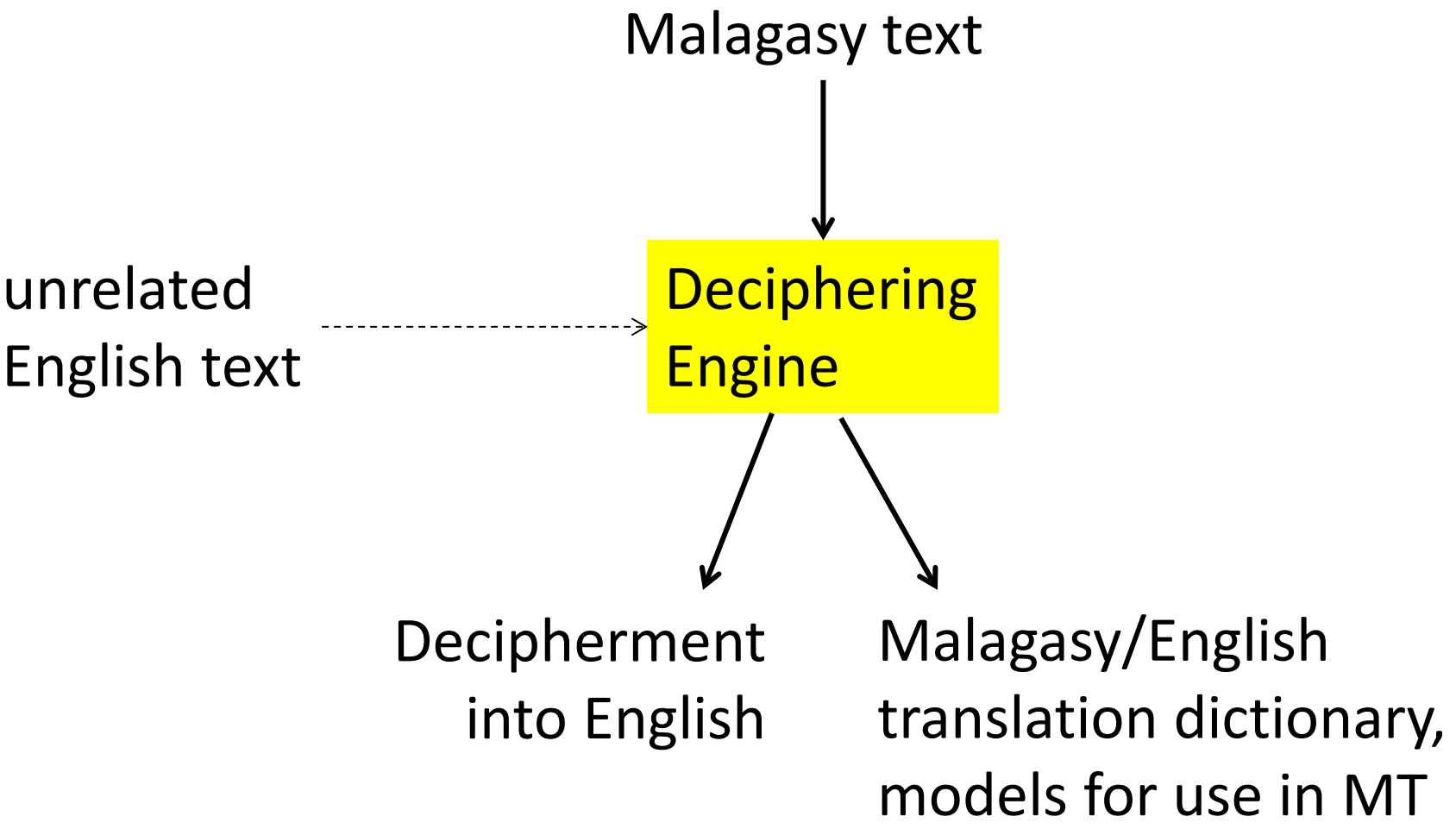
Foreign La
as a Code fo

<n> Hzyr!n <n>

13	4 Hzyr!n 1967
12	fy 12 Hzyr!n 1993
7	5 Hzyr!n 1967
6	fy 30 Hzyr!n 1989
6	30 Hzyr!n 1989
4	fy 30 Hzyr!n 1994
4	fy 30 Hzyr!n 1993
3	fy 19 Hzyr!n 1967
2	ywm 30 Hzyr!n 1989
2	w 6 Hzyr!n 1994
2	qbl 5 Hzyr!n 1967
2	fy 9 Hzyr!n 1967
2	fy 7 Hzyr!n 1981
2	fy 6 Hzyr!n 1994
2	fy 5 Hzyr!n 1967

Search query	Documents
January 4, 1967	8040
February 4, 1967	9270
March 4, 1967	10700
April 4, 1967	21800
May 4, 1967	14000
June 4, 1967	39300
July 4, 1967	12600
August 4, 1967	7970
September 4, 1967	7390
October 4, 1967	8800
November 4, 1967	6560
December 4, 1967	9770

Exploit Non-Parallel Data for MT



Deciphering Malagasy into English

20m word tokens

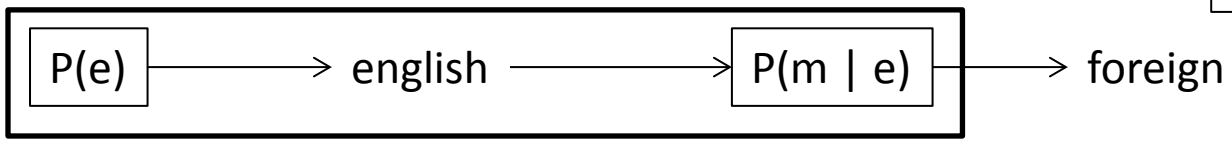
foreign

Deciphering Malagasy into English

learned on unrelated english

learned, to maximize P(foreign-corpus)

20m word tokens



(language model)

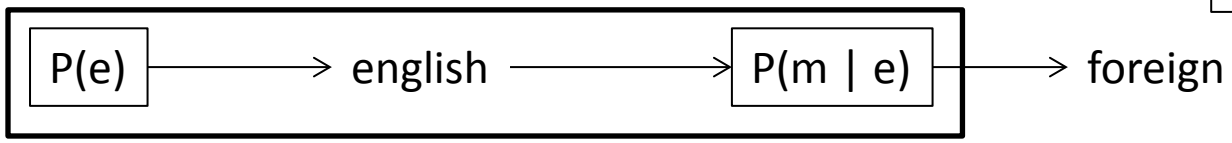
(bilingual translation dictionary, re-ordering patterns, etc)

Deciphering Malagasy into English

learned on unrelated english

learned, to maximize P(foreign-corpus)

20m word tokens



(language model)

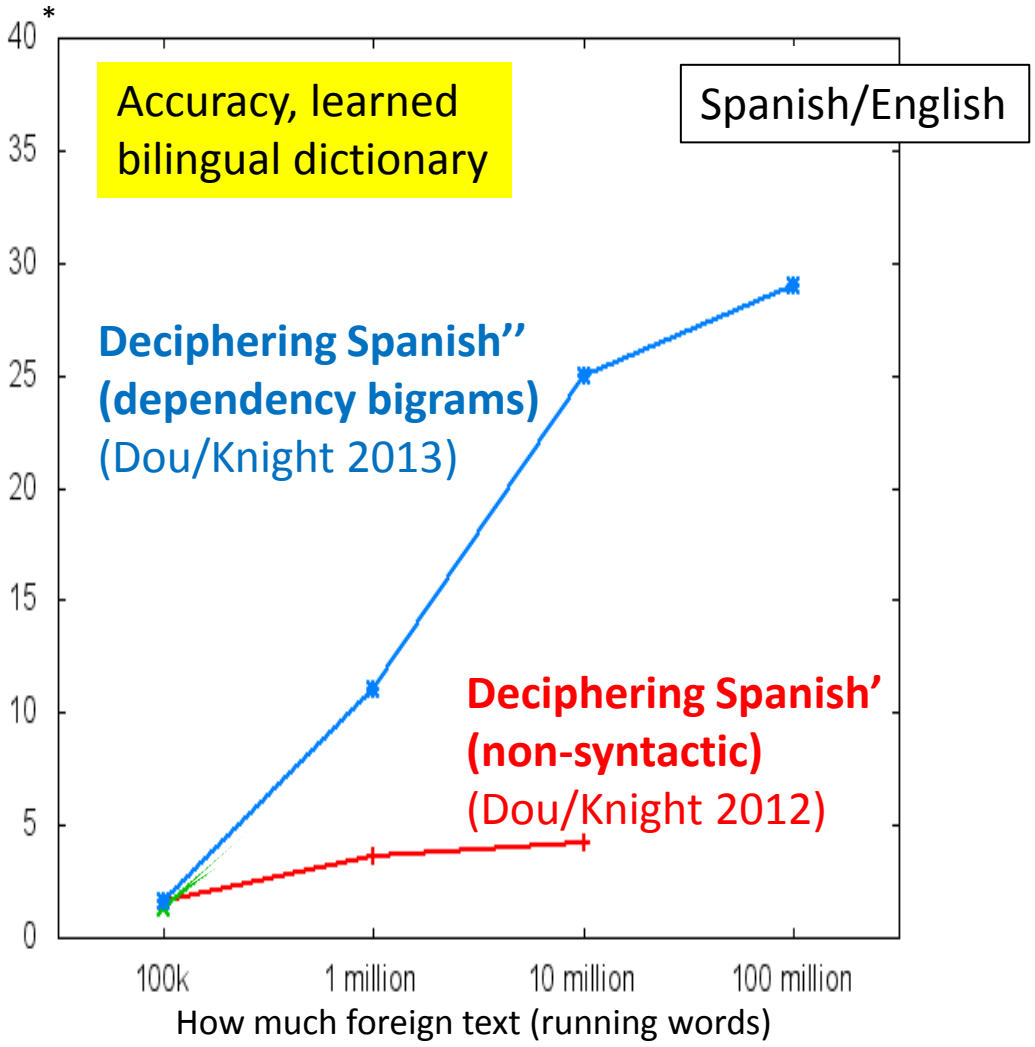
(bilingual translation dictionary, re-ordering patterns, etc)

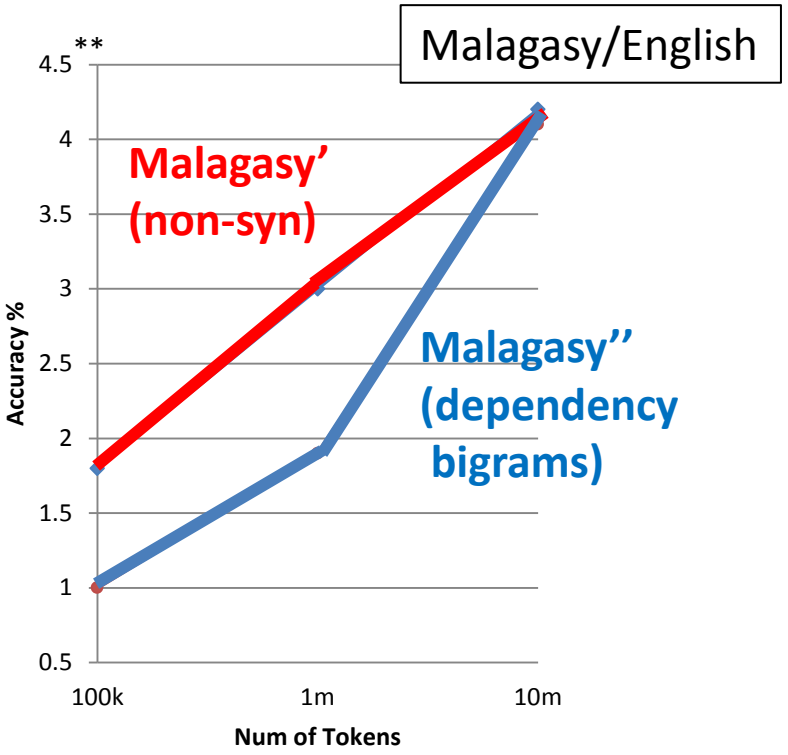
for example

Malagasy	English
maro	many
monisipaly	municipal
ratsy	bad
midadasika	large
vavy	female
lalina	fundamental
manokana	special
taitra	surprised

learned from non-parallel text by our method

Decipherment Accuracy Needs Good Dependency Parsing





Low-quality dependencies hurt decipherment, instead of helping!

* of most freq 5000 word types, 1-best translation in parallel dict

** of most freq 5000 word types, any of 5-best in parallel dict

Malagasy Dependency Parsing

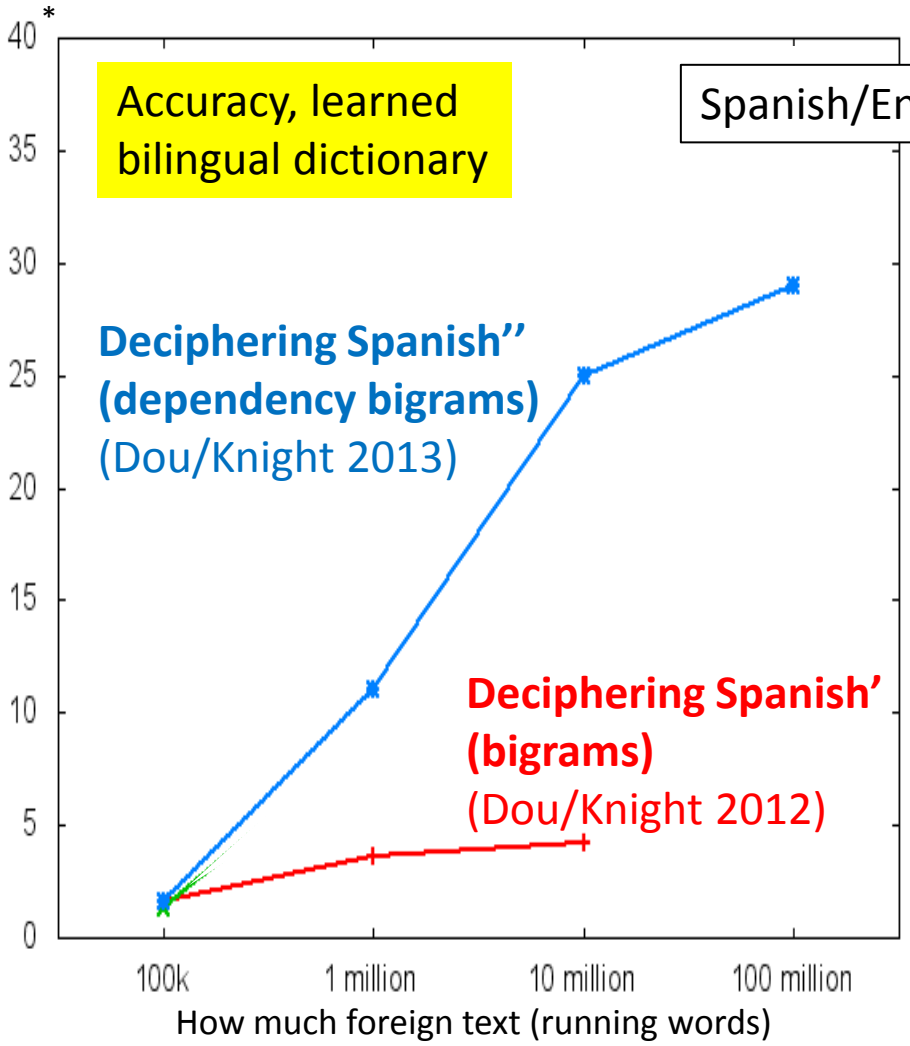
# of sentences	source	manual POS tags?	manual treebank?	parallel English?	manual alignment?
168 (MIT)	IGT, Lagazette, Lakroa	yes	yes		
620 (UT)	GlobalVoices	yes		yes	

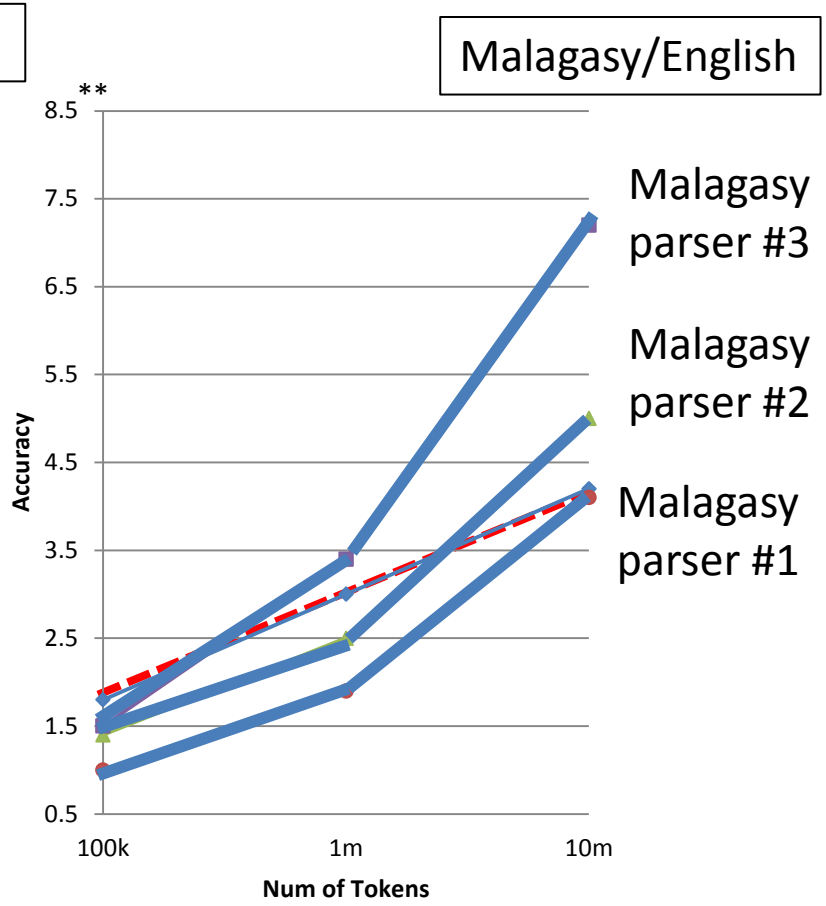
Malagasy Parser #1
Train CMU Turbotagger and Turboparser on 168 sentences (up from 120 in CMU version)

Malagasy Parser #2
Manually fix POS tags with help of **online dictionary interface** and **parallel data**
Train UT tagger on 168 + 620 = 788 sentences
Train CMU Turboparser on 168 trees **with automatic POS tags from UT tagger**

Malagasy Parser #3
Extend treebank by automatically parsing 246 GlobalVoices sentences,
Manually correct POS tags and dependency links
Train UT tagger on 168 + 620 + 246 = 1034 sentences (up from 788)
Train CMU Turboparser on 168 + 246 = 414 sentences (up from 168)

Decipherment Accuracy Needs Good Dependency Parsing





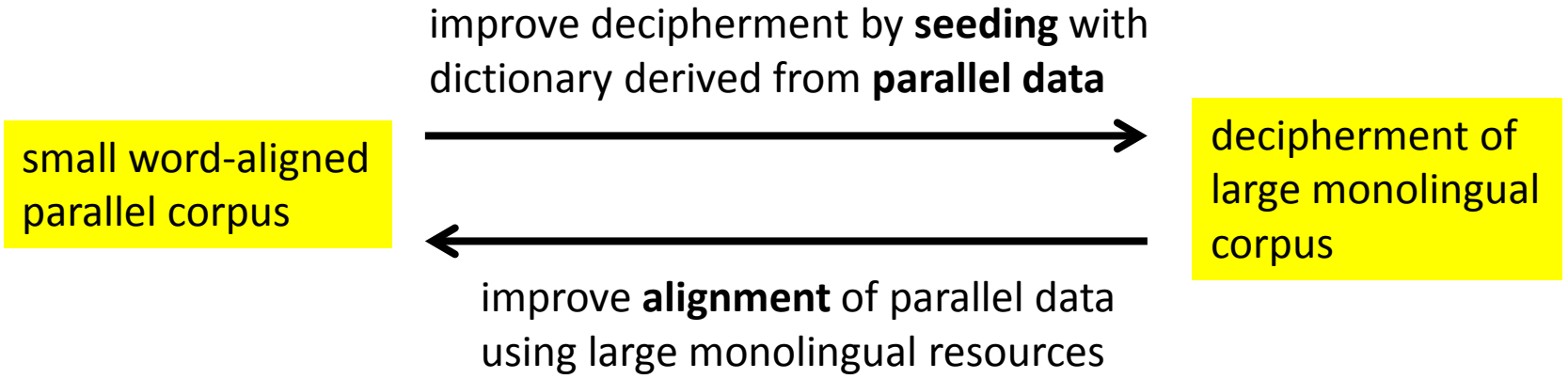
* of most freq 5000 word types, 1-best translation in parallel dict

** of most freq 5000 word types, any of 5-best in parallel dict

Next Steps

Use decipherment results to improve Malagasy-to-English MT

Virtuous improvement cycle:



end