From SMS Gathering to SMS Normalization

Linguistic Studies and Finite-State Algorithms

Richard Beaufort

UCL
Université
catholique
de Louvain

Contents

Contents

SMS ?

- SMS = Short Message Service
- Written messages between mobile phones
- Introduced a few years ago
- Quickly adopted by users
- Deviations from traditional spelling conventions
  - But what exactly ?
  - « SMS » language ?
  - Aim: corpus-based linguistic studies

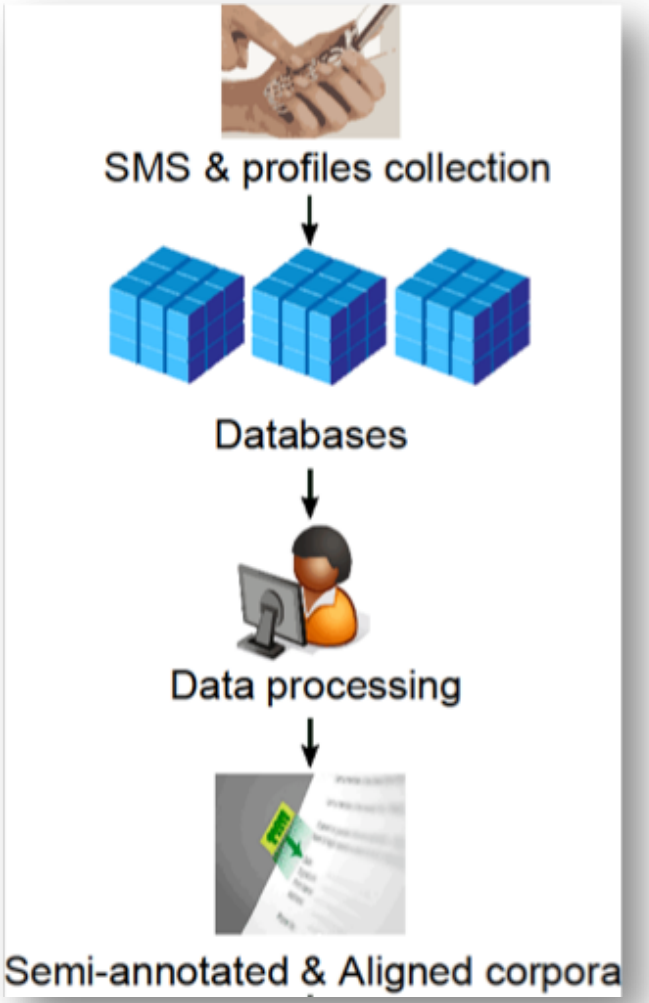
Contents

Projects

- 2 successive projects
  - Faites don de vos SMS à la Science (2004-2005)
    - Belgian French
  - sms4science (International, 2007-2011)
    - **Cf. Amélie Cougnon** (louise-amelie.cougnon@uclouvain.be)
    - Other variants of French:
      - La Réunion, Switzerland and Québec
      - Coming soon: France
    - Other languages:
      - Italian and German (Switzerland)
      - Coming soon: English (Québec, UK), Spanish, Greek.
      - What about Dutch and Flemish ?

Gathering

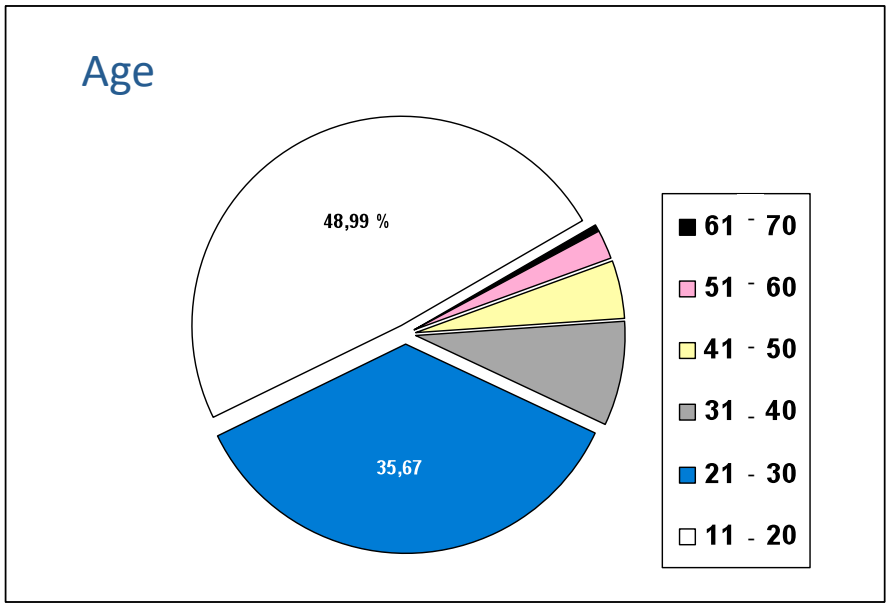
Proximus
70.000 messages
2.773 profiles

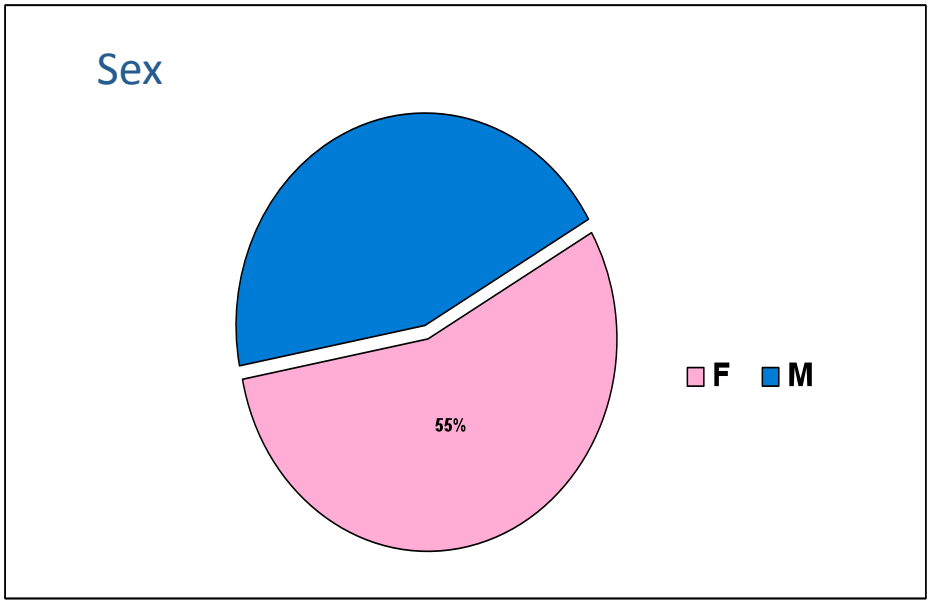


30.000 messages
- Anonymized
- Transcribed
- Annotated

Objective ?

- Not completely...
  - not random
  - volunteers
    - chose what they sent...



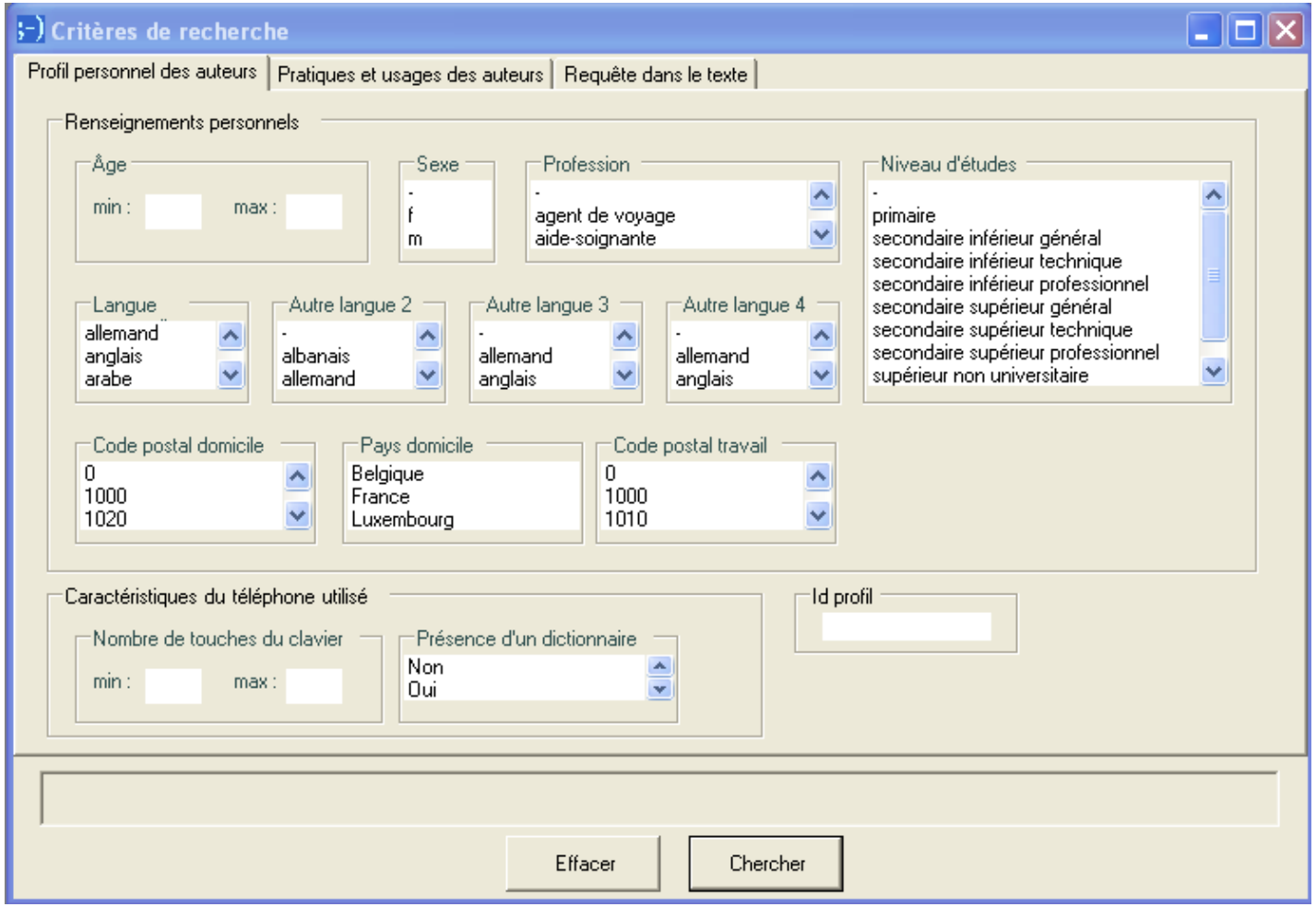


Contents

1. SMS ?
2. SMS gathering
3. The SMS Language
   - Tool
   - Abbreviations
   - Greetings
   - Loan words
4. SMS normalization

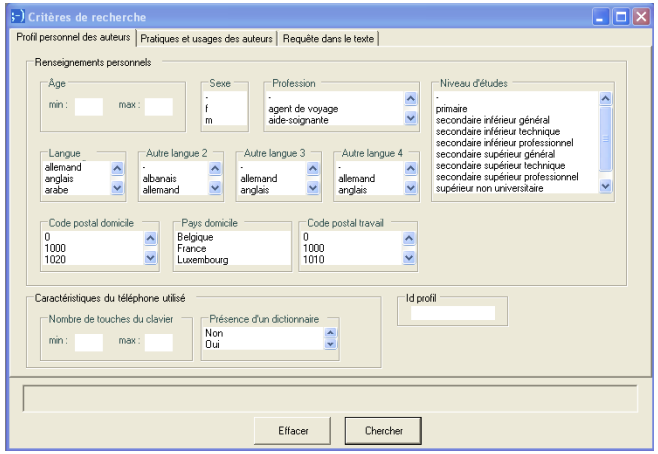
Tool

- Dedicated interface



Tool

- Dedicated interface

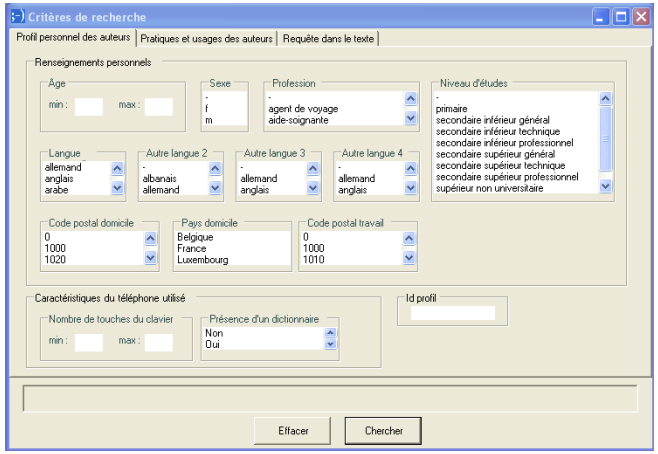


Statistics

- Abbreviations, word frequencies, loan words, etc.
- Age, sex, education, etc.

Tool

- Dedicated interface



Statistics

- Abbreviations, word frequencies, loan words, etc.
- Age, sex, education, etc.

Unitex

- Regular expressions
- UMLV, Paris-Est, France
- http://www-igm.univ-mlv.fr/~unitex/

Statistics

- In next slides, all statistics about:
  - Abbreviations
  - Greetings
  - Loan words

  come from:
  - Cougnon L.-A. et François T., Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS in *Proceedings of 10th International Conference JADT*, 9-11 juin 2010, Sapienza University of Rome, volume 1.
  - Cf. http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0619-0630_036-Cougnon.pdf

Statistics

- Warning: results must be seen as
  - general **trends**
  - coming from a corpus that we showed as not necessarily representative (cf. previous slide about « objectivity of the data »)

Statistics - Abbreviations

« Eseydepa C àlamésonce WE »

Statistics - Abbreviations

« Eseydepa C àlamésonce WE »

(24 characters)

« Essaie de passer à la maison ce week-end »

[*Try and come home this weekend*]

(40 characters)

40% shorter !

Statistics - Abbreviations

« Eseydepa C àlamésonce WE »

- Word boundaries
  - Eseydepa C
  - Essaie de passer
- Letter changes (esey – essaie [ɛ s ɛ j])
- Phonetic letters (c – [pa]sser [s e])
- abbreviations (WE – weekend)

Statistics - Abbreviations

- Phonetic numbers
  - 2m1 (**de**m**ain** [d ə m ɛ̃])
  - b1 (b**ien** [bjɛ̃])
  - v1 (v**iens**, v**ient** [vjɛ̃])
  - v2 (v**eux**, v**eut**, v**œux** [v ø], v**ieux** [v j ø])
- Consonant skeleton
  - bjr (bonjour [b ɔ̃ ʒ u ʁ])
  - slt (salut [s a l y])

Statistics - Abbreviations

- 105 characters / transcription
  - SMS = 9.4% shorter (13.4 characters)
- All messages? No
  - « Only » 78.6%
  - Same size: 19%
  - Longer: 2.4% (aaaaaahhhh !)

Statistics - Abbreviations

- Sex:
  - Women: 10% (109 characters)
  - Men: 8.4% (99 characters)
- Age:
  - < 15: 15.4%
  - > 45: 4.5%

Statistics - Abbreviations

- Education
  - Primary (6 – 12) and secondary, part 1 (12-15)
    - 13.9%
  - Secondary, part 2 (15-18)
    - general: 11%
    - technical: 9.5%
  - post-secondary education
    - Trade schools: 7%
    - University: 8.5%

Statistics - Greetings

- Standard registers
  - **bonjour**: neutral
  - salut: casual
  - coucou: familiar
  - kikou: very informal, almost slang
- Note:
  - In next slides, we present the most specific form (not necessarily the most frequent) for each group
  - Method: chi-square

Statistics - Greetings

- Sex
  - Women: coucou
  - Men: salut, **bonjour**
- Age
  - < 15 : salut
  - 15-19 : kikou
  - 20-24: coucou
  - > 25: **bonjour**
  - 25-34: **hello**, **hi**
  - > 45: coucou

Statistics - Greetings

- Education
  - Primary (6 – 12) and secondary, part 1 (12-15)
    - salut
  - Secondary, part 2 (15-18)
    - general: salut
    - technical: **bonjour**
  - post-secondary education (18+)
    - Trade schools: **bonjour**
    - University: coucou, **hello**, **hi**

Statistics - Loan words

- 536 distinct words
- In 9.5% messages
- 16 languages: English, Italian, Dutch, Arabic, Spanish...
- > 100 occ: English only
  - open/close message (hello, bye)
  - shorter, easier (today >< aujourd'hui)
- Italian: love

Statistics - Loan words

- Age: especially < 25
- Education
  - especially:
    - Primary (6 – 12) and secondary, part 1 (12-15)
    - Technical secondary (15-18)
  - almost never
    - Trade school
- Words from known language: 30%
  - Maybe environment ?

Contents

1. SMS ?
2. SMS gathering
3. The SMS Language
4. SMS normalization
   - *What is SMS normalization?*
   - *Interest of SMS normalization?*
   - *Whatever the system...*
   - **Main features of our system**
   - **Originality of our system**
   - **Evaluation**
   - **Future works**

What is SMS normalization ?

- Rewrite/transcribe an SMS
  using a more conventional spelling (><noisy)

English

SMS: btw wenz ur flt 2moro

French

SMS: slt cmt cv? b1? mwa c ok. keskon fé?

What is SMS normalization ?

- Rewrite/transcribe an SMS
  using a more conventional spelling (><noisy)

English

SMS: btw wenz ur flt 2moro

Transcription: By the way, when is your flight tomorrow?

French

SMS: slt cmt cv? b1? mwa c ok. keskon fé?

What is SMS normalization ?

- Rewrite/transcribe an SMS
  using a more conventional spelling (><noisy)

English

SMS: btw wenz ur flt 2moro

Transcription: By the way, when is your flight tomorrow?

French

SMS: slt cmt cv? b1? mwa c ok. keskon fé?

Transcription: Salut, comment ça va? Bien? Moi, c'est ok. Qu'est-ce qu'on fait?

Interest of SMS normalization ?

- Makes SMS more readable…
  - **for machines**
    - TTS synthesis engines (analyzes texts…)

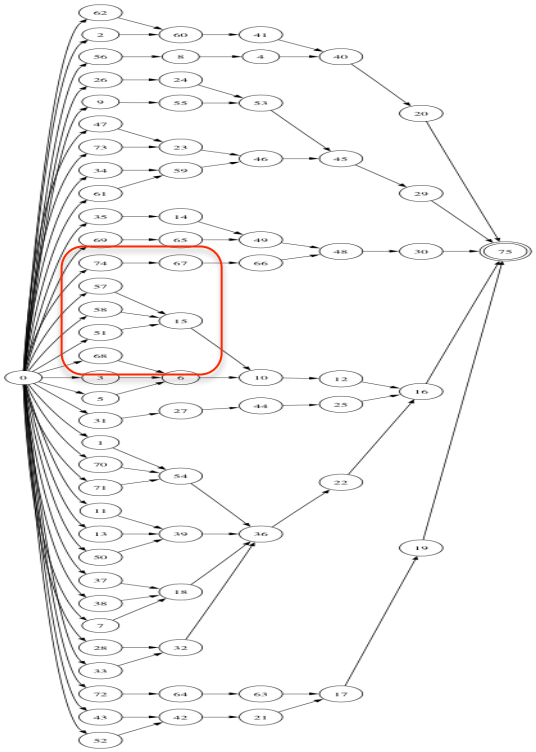
      « slt cmt cv? » « Salut, comment ça va? »
    - Information retrieval tools (conventional forms)
  - **for human beings**
    - not all used to decipher the SMS language

Whatever the system…

- **A way of rewriting** « noisy » SMS sequences
  - « rewrite rules » (unweighted/weighted)
    - C ⟶ R :: G _ D / w
    - « n » ⟶ « m » :: <V> _ (p|b) (enporter ⟶ emporter)
  - different interpretations
    - spell checker (spell error detected)
    - machine translation (phrase-based transcription)
    - auto. speech recognition (phonetic « encryption »)

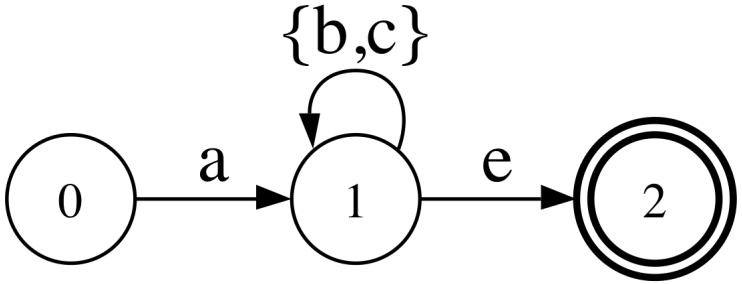
Whatever the system…

- **A way of choosing** between solutions
  - context (words, syntax)



Main features of our system

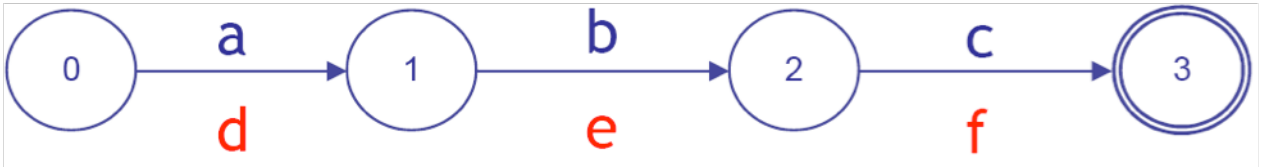
- Finite-state machines
  - automaton



  - regular language
  - regular expression: a (b|c)* e

Main features of our system

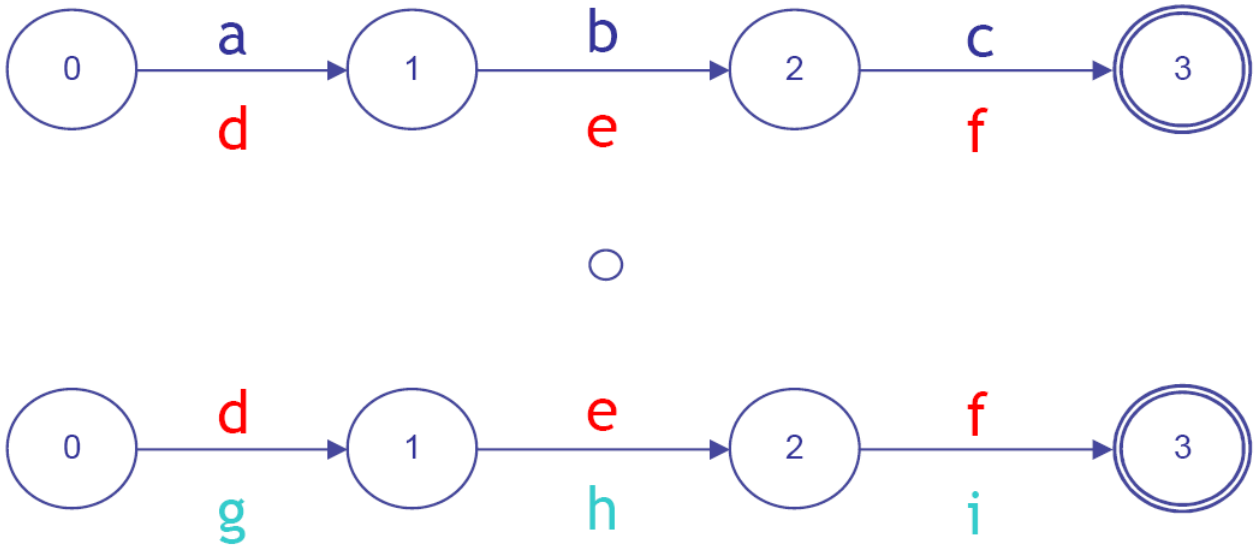
- Finite-state machines
  - transducer



- relations
- rewrite rules: abc ⟶ def

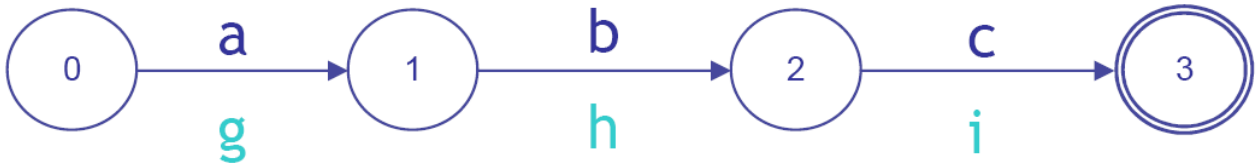
Main features of our system

- Finite-state machines
  - transducers: composition



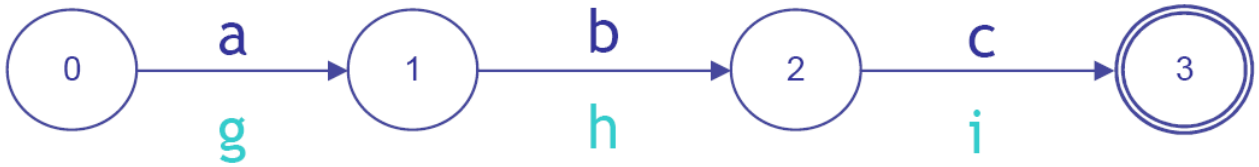
Main features of our system

- Finite-state machines
  - transducers: composition



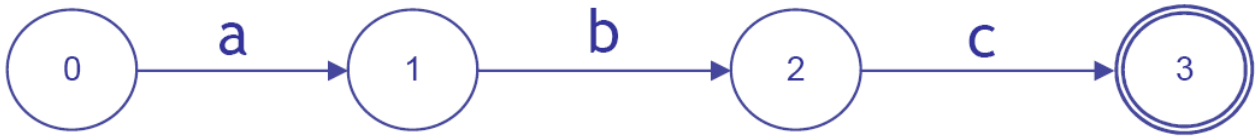
Main features of our system

- Finite-state machines
  - transducer: projections



Main features of our system

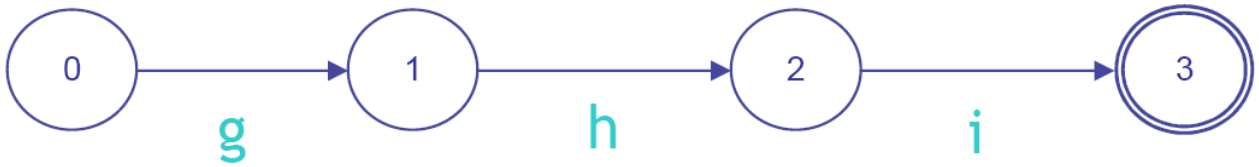
- Finite-state machines
  - transducer: projections



    - first projection

Main features of our system

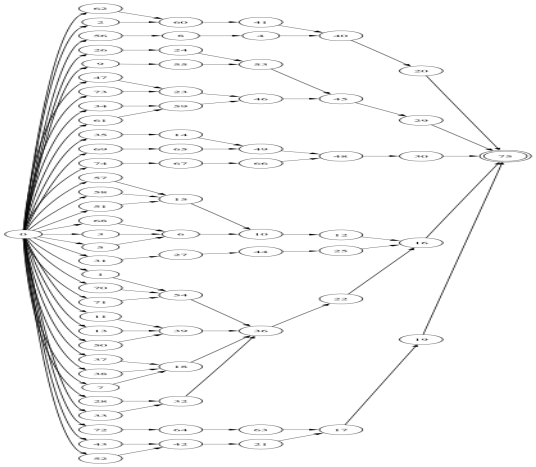
- Finite-state machines
  - transducer: projections



    - second projection

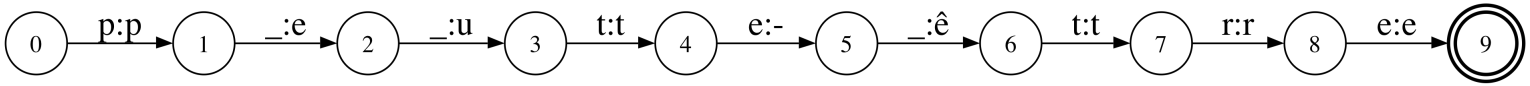
Main features of our system

- Finite-state machines
  - best path



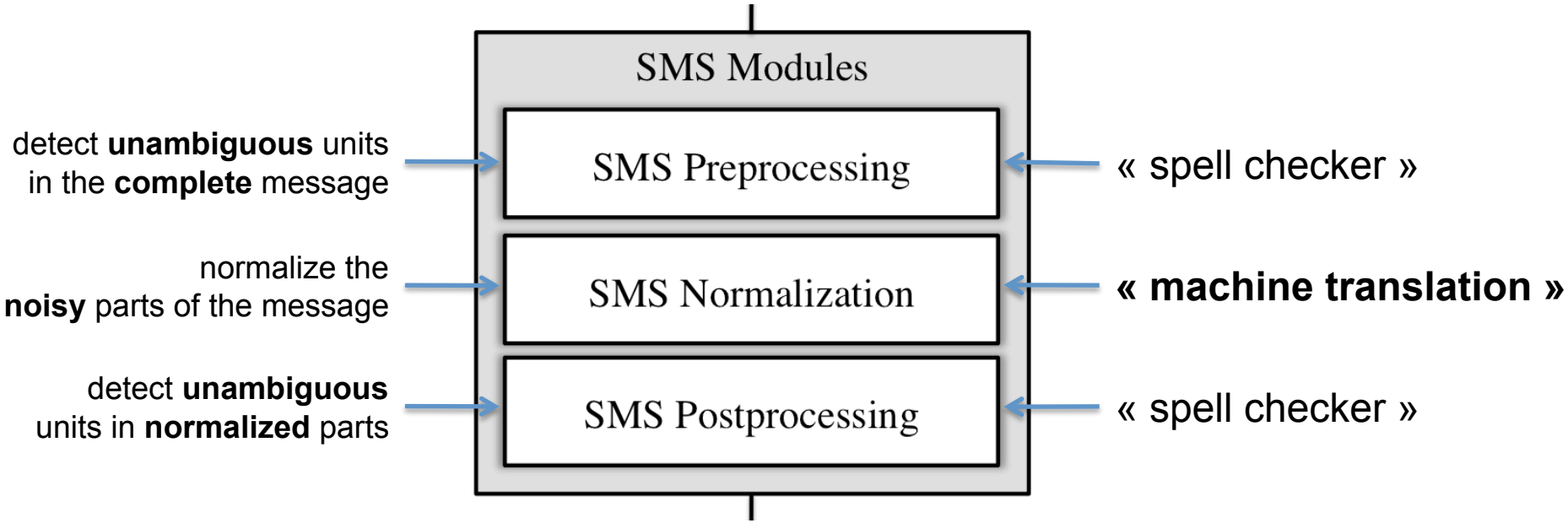
Main features of our system

- Finite-state machines
  - best path



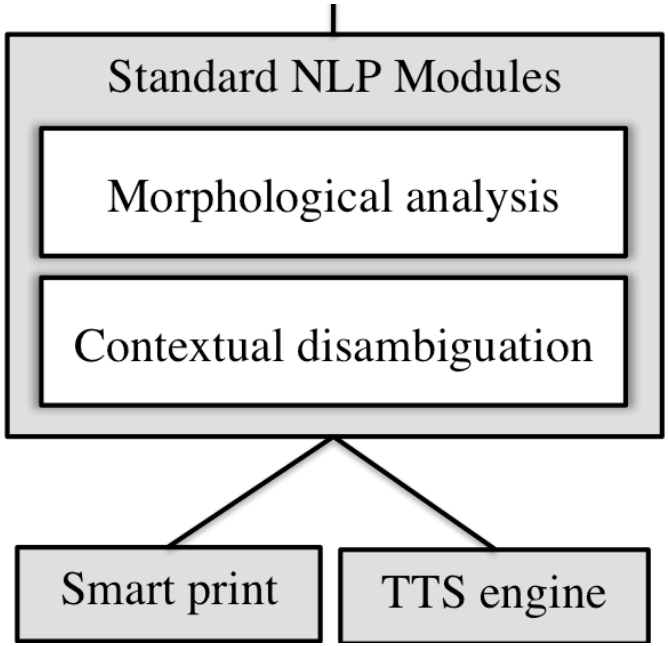
Main features of our system

- Finite-state machines
- Combines spell checker/machine translation



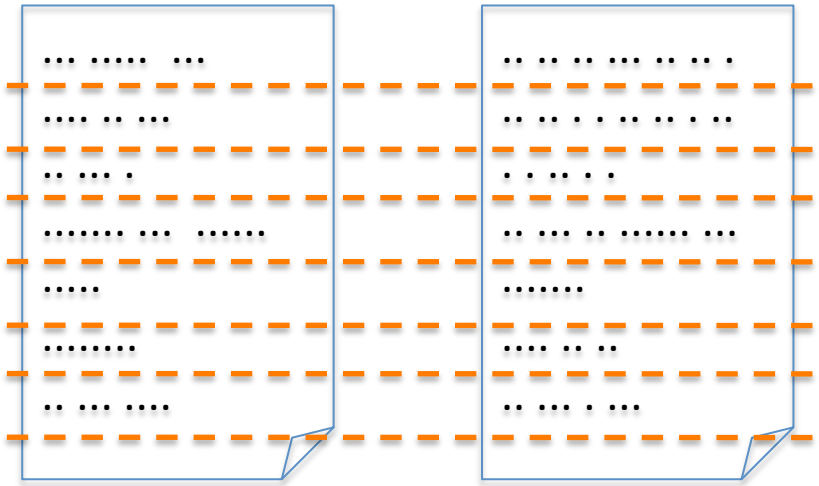
Main features of our system

- Finite-state machines
- Combines spell checker/machine translation
- Standard natural language processing included



Main features of our system

- Finite-state machines
- Combines spell checker/machine translation
- Standard natural language processing included
- Normalization learned from 2 parallel corpora



Originality of our system

- Normalization learned from parallel corpora

```
j esper ktt vb1...jpenses atwa
J'espère que tout va bien... Je pense à toi
```

Originality of our system

- Normalization learned from parallel corpora

```
j esper ktt vb1...jpenses atwa
J'espère que tout va bien... Je pense à toi
```

```
j esper_ k___t__t v__b1__..._j__penses a_twa
J'espère que tout va bien... Je pense_ à toi
```

Originality of our system

- Normalization learned from parallel corpora

GIZA++ ?

- Och & Ney (2003)
- http://fjoch.com/GIZA++.html
- Machine Translation
- « only » Word alignment

Originality of our system

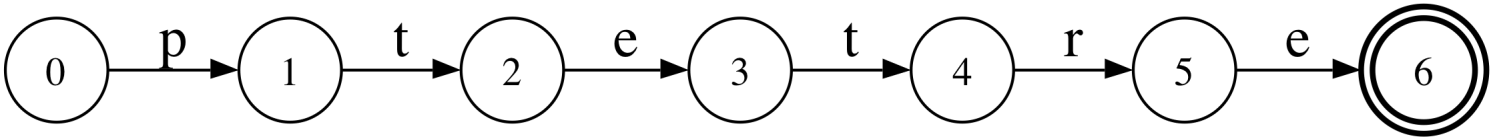
- Normalization learned from parallel corpora

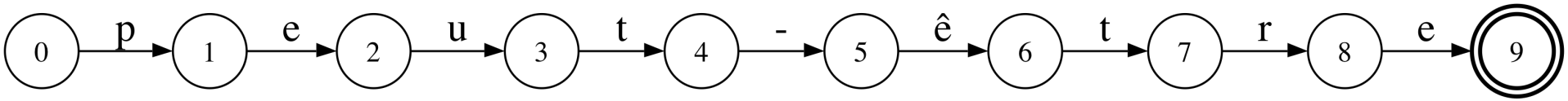
p t e t r e

p e u t - ê t r e

Originality of our system

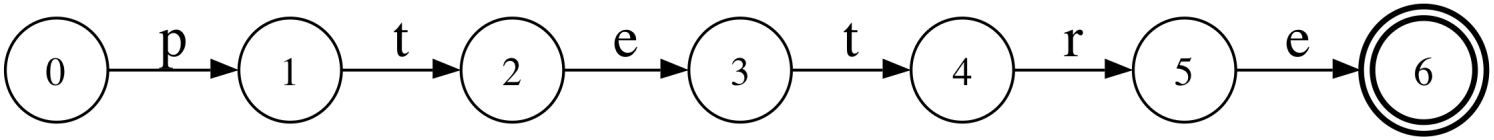
- Normalization learned from parallel corpora



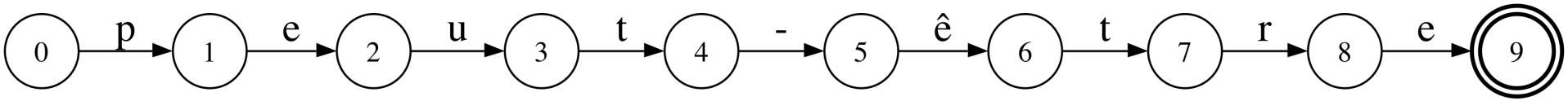


Originality of our system

- Normalization learned from parallel corpora

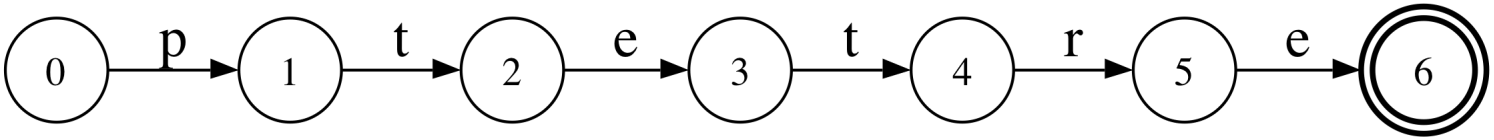


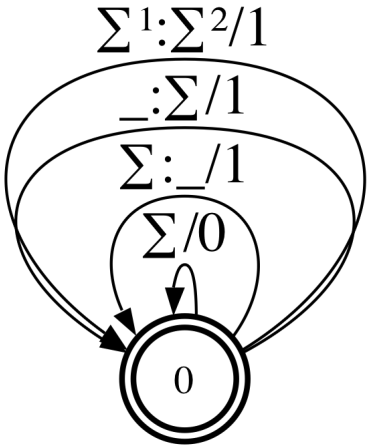
idem:	0
substitution:	1
insertion:	1
deletion:	1

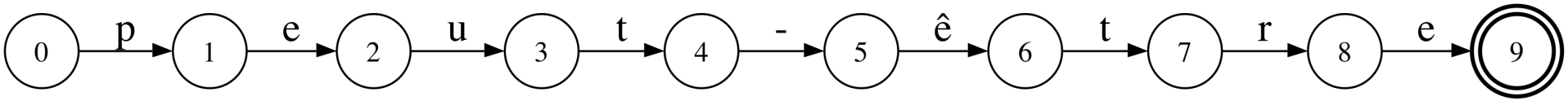


Originality of our system

- Normalization learned from parallel corpora

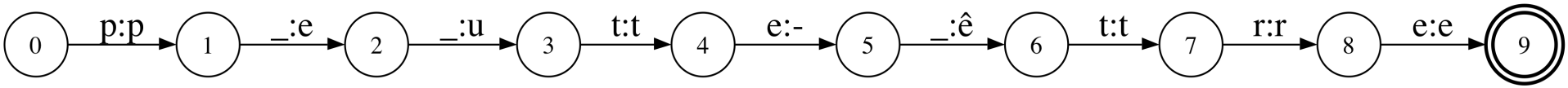






Originality of our system

- Normalization learned from parallel corpora

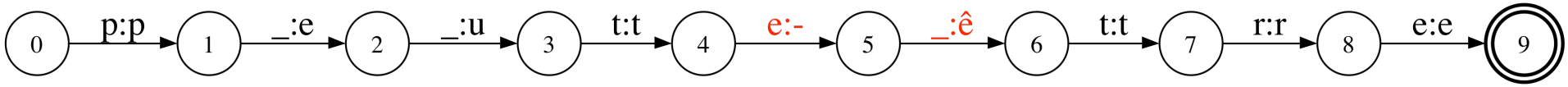



```
p _ _ t e _ t r e

p e u t - ê t r e
```

Originality of our system

- Normalization learned from parallel corpora



p _ _ t e _ t r e

p e u t _ ê t r e

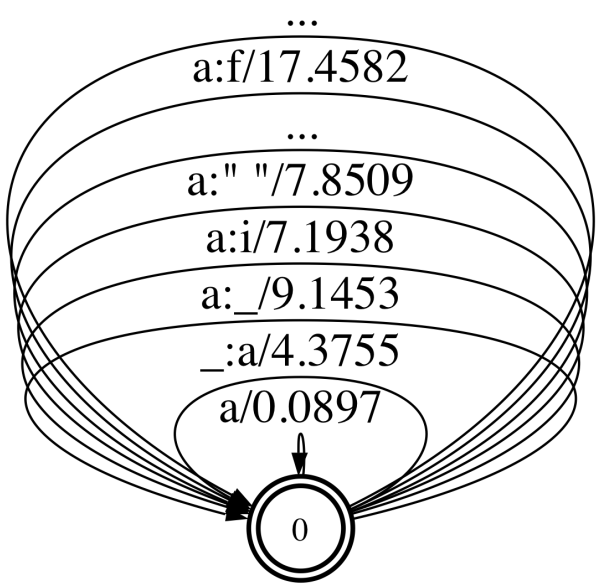
Originality of our system

- Normalization learned from parallel corpora

id(a)	/	0.0897
ins(a)	/	4.3755
del(a)	/	9.1453
sub(a,i)	/	7.1938
sub(a,« »)	/	7.8509
...		
sub(a,f)	/	17.4582
...		

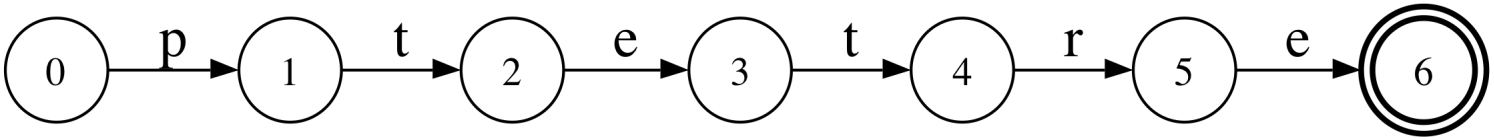
Originality of our system

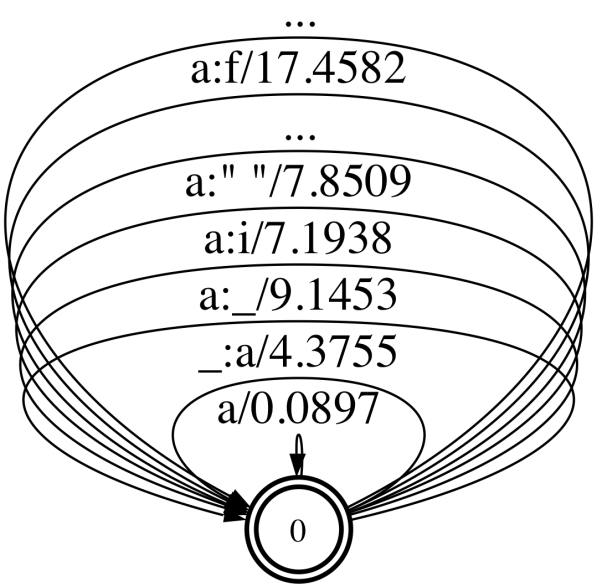
- Normalization learned from parallel corpora

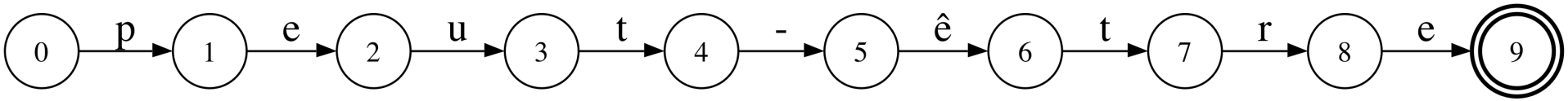


Originality of our system

- Normalization learned from parallel corpora

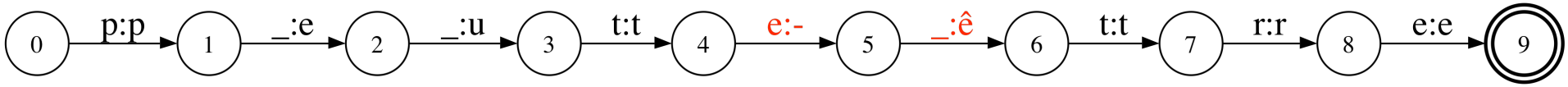






Originality of our system

- Normalization learned from parallel corpora



p _ _ t e _ t r e

p e u t _ ê t r e

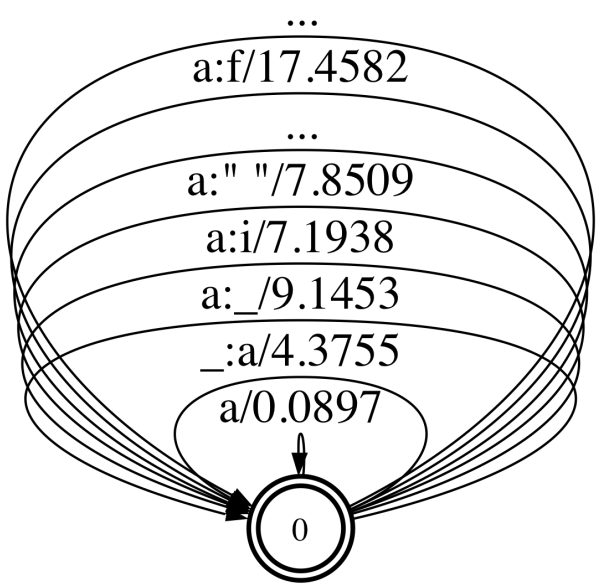
Originality of our system

- Normalization learned from parallel corpora


```
p _ _ t _ e t r e
      ----
p e u t - ê t r e
      ----
```

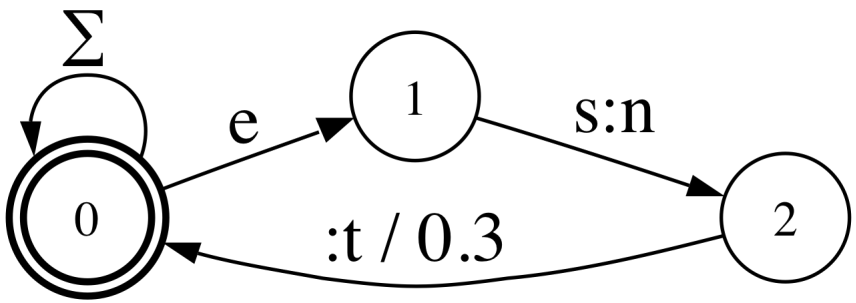
Originality of our system

- Normalization learned from parallel corpora



Originality of our system

- Normalization learned from parallel corpora



Originality of our system

- Normalization learned from parallel corpora

es ⟶ ent / 0.3

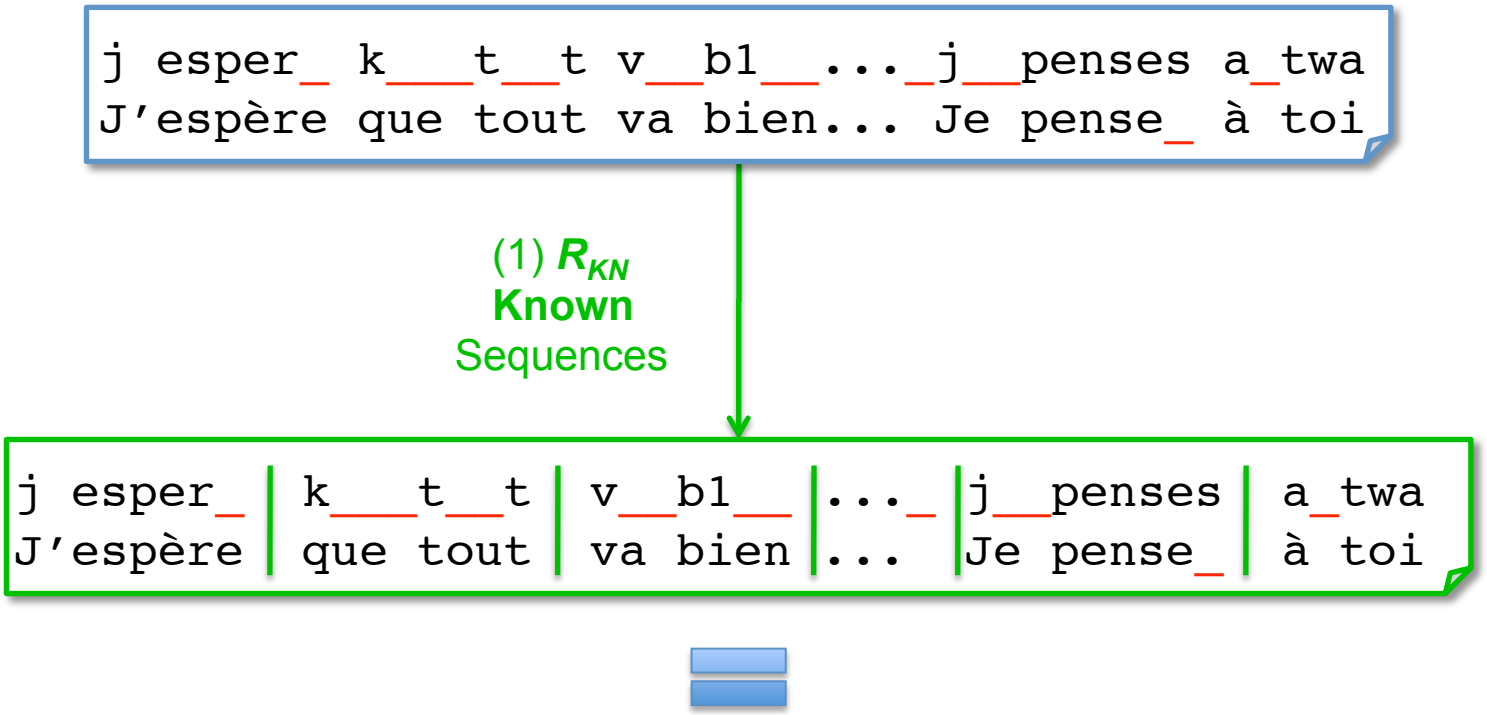
Originality of our system

- Normalization learned from parallel corpora

```
j esper_ k___t__t v__b1__..._j__penses a_twa
J'espère que tout va bien... Je pense_ à toi
```

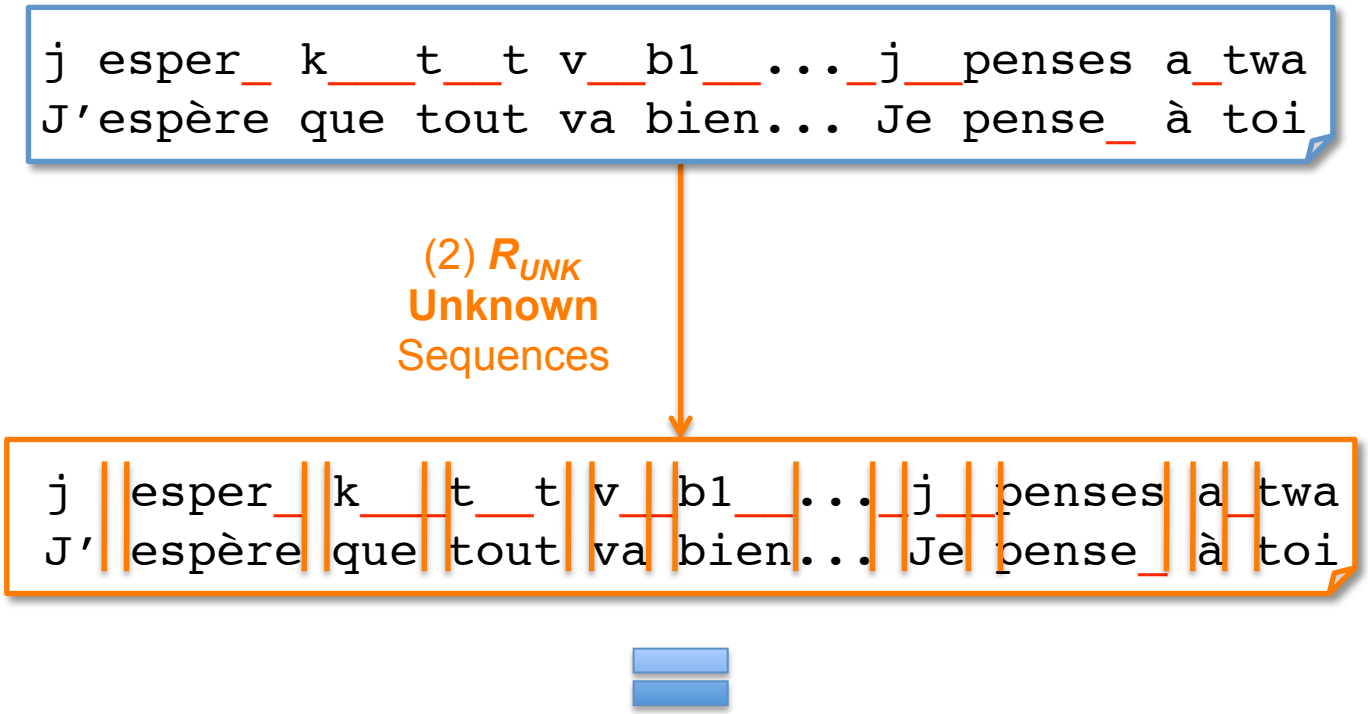
Originality of our system

- Normalization learned from parallel corpora



Originality of our system

- Normalization learned from parallel corpora



(j), (esper), (k), (tt), (v), (b1), (j), (penses), (a), (twa)
(J'), (espère), (que), (tout), (va), (bien), (je), (pense), (à), (toi)

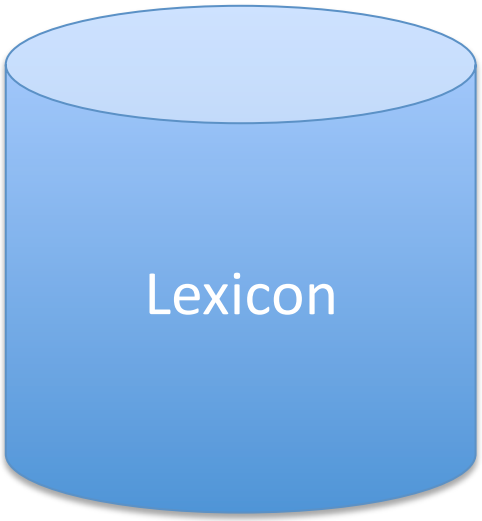
Originality of our system

- Normalization learned from parallel corpora

```
j esper_ k___t__t v__b1__..._j__penses a_twa
J'espère que tout va bien... Je pense_ à toi
```

- ktt → que tout
- kt → que tou
  kt
- tt → tout
  tt

look-up



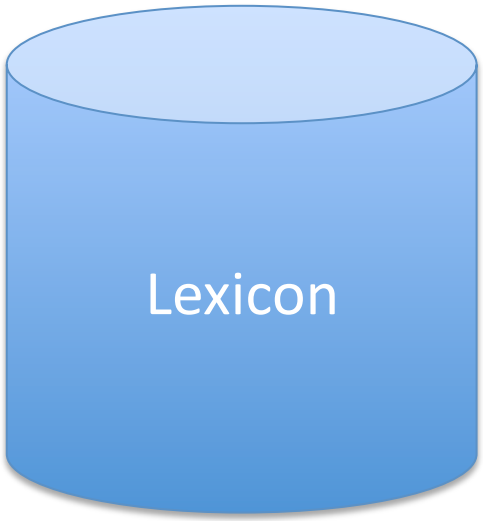
Originality of our system

- Normalization learned from parallel corpora

```
j esper_ k___t__t v__b1__..._j__penses a_twa
J'espère que tout va bien... Je pense_ à toi
```

- ktt ⟶ que tout
- kt ⟶ que tou
  kt
- tt ⟶ tout
  tt

look-up ⟷



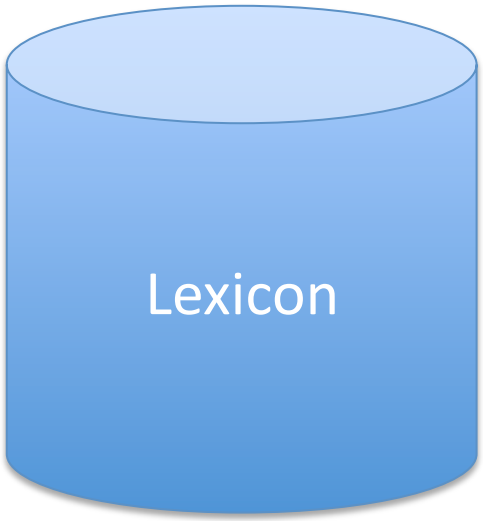
Originality of our system

- Normalization learned from parallel corpora

```
j esper_ k___t__t v__b1__..._j__penses a_twa
J'espère que tout va bien... Je pense_ à toi
```

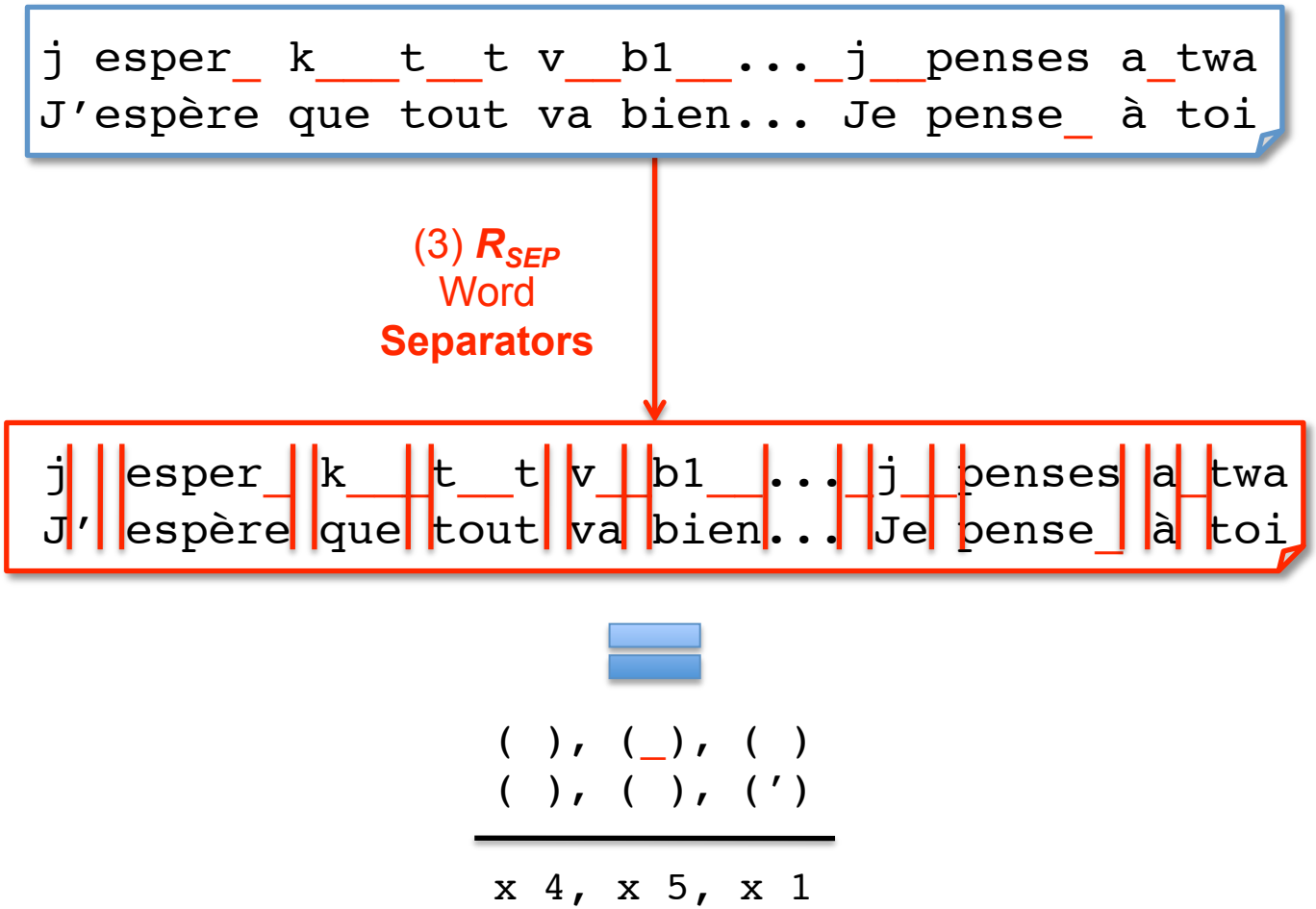
- ktt → que tout
- tt → tout
  tt

look-up



Originality of our system

- Normalization learned from parallel corpora



Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$\{S_0, S_1, \ldots, S_{n-1}, S_n\} = S \circ Split$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$\{S_0, S_1, \ldots, S_{n-1}, S_n\} = S \circ Split$

$Split = [Sep^* \ (Kn \mid Unk) \ (Sep^+ \ (Kn \mid Unk))^* \ Sep^*]$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$\{S_0, S_1, \ldots, S_{n-1}, S_n\} = S \circ Split$

$Split = [Sep^* \ (Kn \mid Unk) \ (Sep^+ \ (Kn \mid Unk))^* \ Sep^*]$

$Kn = \text{Proj}_1(R_{KN})$

$Sep = (\ \text{Proj}_1(R_{SEP})\ )^*$

$Unk = (\ \text{Compl}(Kn) \cap \text{Compl}(.^*Sep.^*)\ )^*$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6\}$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6\}$

$S_0' = S_0 \circ R_{KN}$
$S_1' = S_1 \circ R_{SEP}$
$S_2' = S_2 \circ R_{UKN}$
$S_3' = S_3 \circ R_{SEP}$
$S_4' = S_4 \circ R_{KN}$
$S_5' = S_5 \circ R_{SEP}$
$S_6' = S_6 \circ R_{UKN}$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6\}$

$$
\begin{aligned}
S_0' &= S_0 \circ R_{KN} \\
S_1' &= S_1 \circ R_{SEP} \\
S_2' &= S_2 \circ R_{UKN} \\
S_3' &= S_3 \circ R_{SEP} \\
S_4' &= S_4 \circ R_{KN} \\
S_5' &= S_5 \circ R_{SEP} \\
S_6' &= S_6 \circ R_{UKN}
\end{aligned}
$$

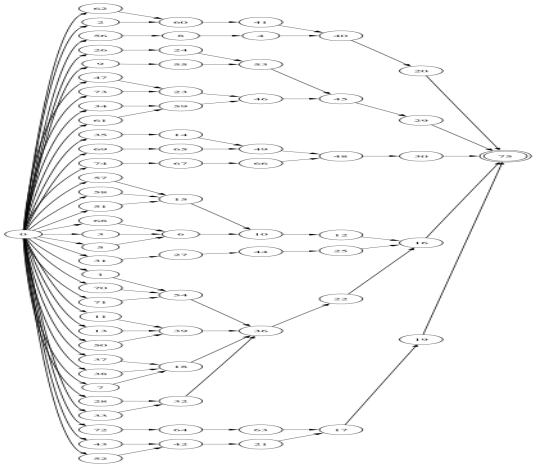
$S' = S_0' \cdot S_1' \cdot S_2' \cdot S_3' \cdot S_4' \cdot S_5' \cdot S_6'$

Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S' = S_0' \cdot S_1' \cdot S_2' \cdot S_3' \cdot S_4' \cdot S_5' \cdot S_6'$



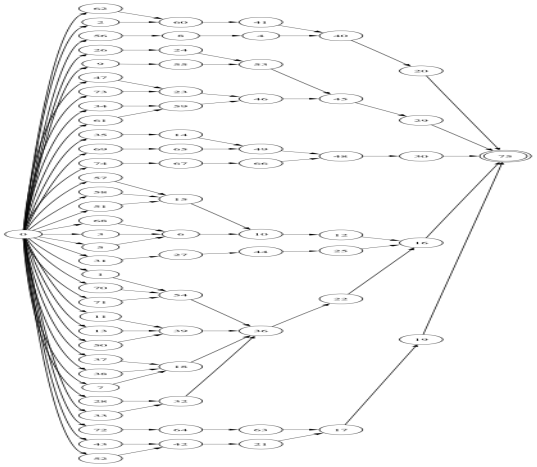
Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S' = S_0' \cdot S_1' \cdot S_2' \cdot S_3' \cdot S_4' \cdot S_5' \cdot S_6'$

$S'' = S' \circ LM$



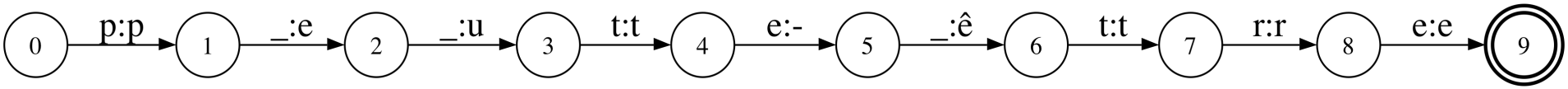
Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S' = S_0' \cdot S_1' \cdot S_2' \cdot S_3' \cdot S_4' \cdot S_5' \cdot S_6'$

$S'' = \text{Best}(S' \circ LM)$



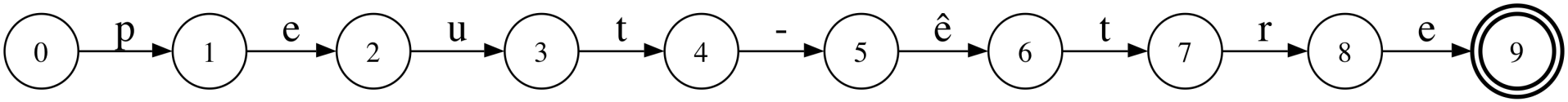
Originality of our system

- Normalization learned from parallel corpora

How do we apply this on a noisy sequence S ?

$S' = S_0' \cdot S_1' \cdot S_2' \cdot S_3' \cdot S_4' \cdot S_5' \cdot S_6'$

$S'' = \text{Proj}_2(\text{Best}(S' \circ LM))$



Evaluation

- Methodology
  - Belgian corpus of 30,000 SMS, 10-fold cross-validation.

Evaluation

- Methodology
  - Belgian corpus of 30,000 SMS, 10-fold cross-validation.
- Efficiency

	mean	**dev.**
bps	1836.57	159.63
ms/SMS (140b)	76.23	22.34

Evaluation

- Methodology
  - Belgian corpus of 30,000 SMS, 10-fold cross-validation.
- Efficiency

	mean	**dev.**
bps	1836.57	159.63
ms/SMS (140b)	76.23	22.34

- Performance

	1. Our approach				2. State of the art					
	Ten-fold cross-validation, French				French			English		
	Copy		*Hybrid*		*Guimier*	*Kobus 2008*		*Aw*	*Choud.*	Cook
	$\bar{x}$	σ	$\bar{x}$	σ	*2007*	1	2*	*2006*	*2006***	*2009***
Sub.	25.90	1.65	6.69	0.45		11.94				
Del.	8.24	0.74	1.89	0.31		2.36				
Ins.	0.46	0.08	0.72	0.10		2.21				
WER	34.59	2.37	**9.31**	0.78		16.51	**10.82**		41.00	44.60
SER	85.74	0.87	**65.07**	1.85		76.05				
BLEU	0.47	0.03	**0.83**	0.01	**0.736**		**0.8**	**0.81**		

$\bar{x}$=*mean*, σ=*standard deviation*

Evaluation

- Analysis
  1. Handles missing and additional word-separators
     - *Pensa ms... -> Pense à mes...*
     - *G T... -> J'étais...*

Evaluation

- Analysis
  1. Handles missing and additional word-separators
     - *Pensa ms... -> Pense à mes...*
     - *G T... -> J'étais...*
  2. Pre- et post- processing useful
     - Unambiguous units not modified

Evaluation

- Analysis
  1. Handles missing and additional word-separators
     - *Pensa ms... -> Pense à mes...*
     - *G T... -> J'étais...*
  2. Pre- et post- processing useful
     - Unambiguous units not modified
  3. Errors often contextual
     - gender: *quel(le)*
     - number: *bisou(s)*
     - person: *tu t'inquiète(s)*
     - tense: *arrivé/arriver*

Future works

1. Phonetic approach
   - Avoid grapheme-to-phoneme conversion at runtime
     - au ⟶[o] ⟶ au, eau, aux, ...
     - Kobus et al. (2008)
   - Learn phonetic similarities
     - Direclty work on graphemic sequences ([o] = o = au = eau = aux = ...)
       - au ⟶ au / 0
       - au ⟶ eau / w

Future works

1. Phonetic approach

2. Another language model
   - SMS corpus: small...
   - Needs:
     - standard written forms
     - but: oral language
   - Solution: blogs
     - « language » oriented
     - ex.: ebooks (http://www.ebooksgratuits.com/)

Future works

1. Phonetic approach

2. Another language model

3. Spell checker inside the NLP modules
   - work on normalized texts
   - contextual errors
   - chart parser

Thank you !

QUESTIONS ?