Language Models for Handwritten Short Message Services

Emmanuel Prochasson, Christian Viard-Gaudin
& Emmanuel Morin

Laboratoire d'Informatique de Nantes Atlantique
Institut de Recherche en Communication et Cybernétique de Nantes
Nantes University

ICDAR 2007

Short Message Services

Features:

- Input with small keypad
- Reduced number of characters allowed
- Fashion amongst teenagers
- New language?
- Spelling liberties

Example:

- *Original text*: Hi mate, Are you okay ? I am sorry that I forgot to call you last night. Why don't we go and see a film tonight ?
- *SMS'ed text*: hi m8 u k ? sry i 4gt 2 cal u lst nyt - y dnt we go c film 2nite

Short Message Services

Features:

- Input with small keypad
- Reduced number of characters allowed
- Fashion amongst teenagers
- New language?
- Spelling liberties

Example:

- *Original text*: Hi mate, Are you okay ? I am sorry that I forgot to call you last night. Why don't we go and see a film tonight ?
- *SMS'ed text*: hi m8 u k ? sry i 4gt 2 cal u lst nyt - y dnt we go c film 2nite

Short Message Services

Features:

- Input with small keypad
- Reduced number of characters allowed
- Fashion amongst teenagers
- New language?
- Spelling liberties

Example:

- *Original text*: Hi mate, Are you okay ? I am sorry that I forgot to call you last night. Why don't we go and see a film tonight ?
- *SMS'ed text*: hi m8 u k ? sry i 4gt 2 cal u lst nyt - y dnt we go c film 2nite

Short Message Services

Features:

- Input with small keypad
- Reduced number of characters allowed
- Fashion amongst teenagers
- New language?
- Spelling liberties

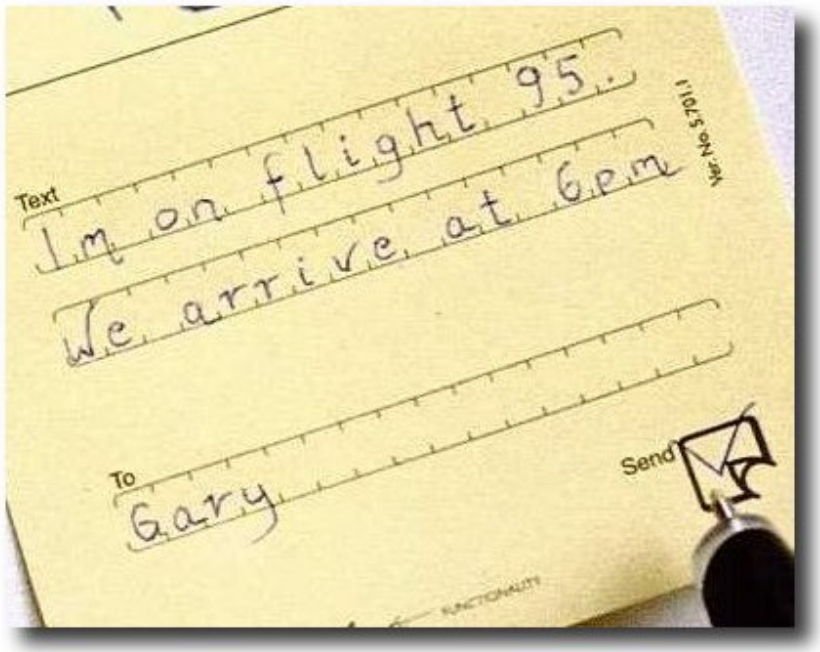
Example:

- *Original text*: Hi mate, Are you okay ? I am sorry that I forgot to call you last night. Why don't we go and see a film tonight ?
- *SMS'ed text*: hi m8 u k ? sry i 4gt 2 cal u lst nyt - y dnt we go c film 2nite

Handwritten Short Message (HSM)

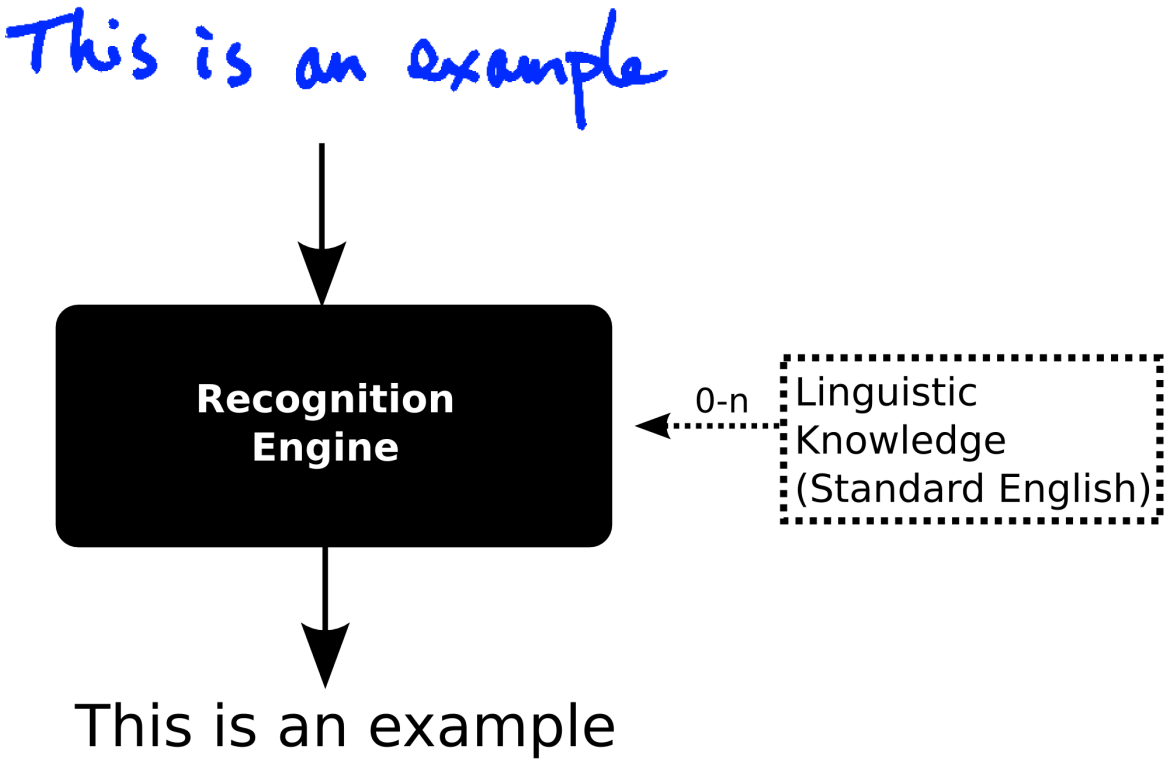
Handwritten Short Message:

- Digital on-line ink input



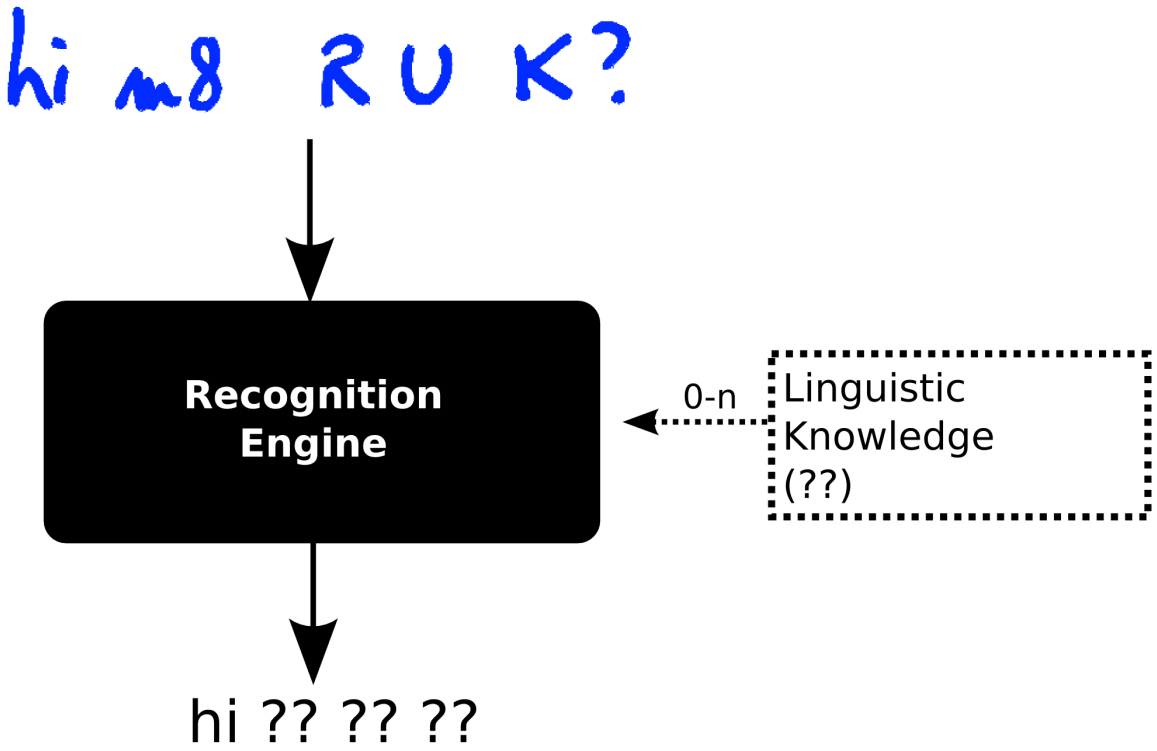
Recognition

Handwriting recognition software



- Very efficient on common text

Handwriting recognition software



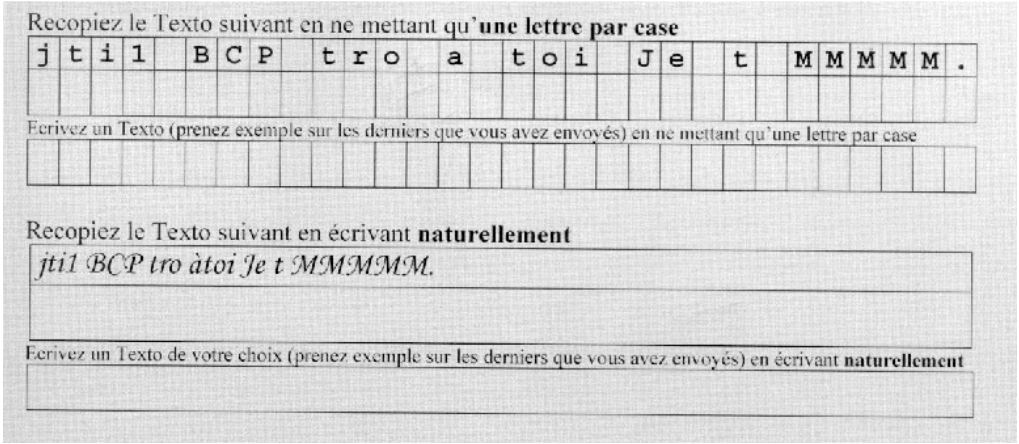
- Unefficient for unknow language and *out-of-lexicon* words

Linguistic Knowledge

- Some LK provided with Handwritting Recognition software (for standard, language-specific text)
- Possibility to create new LK:
  - Generation of a lexicon, in order to cover specific domain vocabulary (example: country name) ;
  - Building a Regular Expression, in order to characterize a form (example: phone number).
- $\Rightarrow$ create new LK to bring HSM-adapted Language Model, in order to assist Handwritting Recognition Process

The HSM Corpora

French corpora collected among student:


Recopiez le Texto suivant en ne mettant qu'**une lettre par case**

j t i 1 B C P t r o a t o i J e t M M M M M .

Ecrivez un Texto (prenez exemple sur les derniers que vous avez envoyés) en ne mettant qu'une lettre par case

Recopiez le Texto suivant en écrivant **naturellement**

jti1 BCP tro àtoi Je t MMMMM.

Ecrivez un Texto de votre choix (prenez exemple sur les derniers que vous avez envoyés) en écrivant **naturellement**

Sample	Boxed	Free
Imposed text	177	174
Non-Imposed text	493	477
Total	670	551
	1321	

More than 38,000 char, 11,600 words

Phenomena separation

- $\longrightarrow$ Original corpora divided depending on each following phenomenon (see Guimier de Neef & Véronis, [2006])
  - **Rebus**: `CU l8er` – *See you later...*
  - **Consonant Skeleton**: `txt` – *text*, `ppl` – *people...*
  - **Phonetic writing**: `giv me som luv` - *give me some love...*
  - + others forms (mostly correct writing).

About Rebus

- Mixing symbols to be read, and symbols to be spelled out ;
- Mixing letters and numbers
- Very creative

examples: `CU l8er` – *See you later*, `2night` – *Tonight*, `X-mas` – *Christmas*...

About Phonetic Writing

- Word (or expression), when read aloud, is *understandable*
  - Examples: `becoz`–*because*, `tonite`–*tonight*
- Lots of possibilities for phonetic writing
- Hard to characterize: no salient morphological clue.

About Phonetic Writing

- Word (or expression), when read aloud, is *understandable*
  - Examples: `becoz`–*because*, `tonite`–*tonight*
- Lots of possibilities for phonetic writing
- Hard to characterize: no salient morphological clue.

About Consonant Skeletons

- Shortening a word by removing most of its vowels: *txt*, *ppl*. . .
- Existing in many languages (English, French, . . .) ;
- Frequently used for long word (ex: `toujours` $\longrightarrow$ `tjrs` – *always*)

Example: Consonant Skeleton

Processed using:

- a lexicon: transformation from a corpora, applying few simple transformation rules...
- ... and a regular expression characterizing the shape of Consonant Skeletons:

Example: Consonant Skeleton

Processed using:

- a lexicon: transformation from a corpora, applying few simple transformation rules...
- ... and a regular expression characterizing the shape of Consonant Skeletons:

Example: Consonant Skeleton

Processed using:

- a lexicon: transformation from a corpora, applying few simple transformation rules...
- ... and a regular expression characterizing the shape of Consonant Skeletons:

Transformation of a word to a Consonant Skeleton

Starting from a word (ex: longtemps – *long* [for a long time]) (Anis [2002])

- vowels at the beginning and at the end are kept → `longtemps`
- other vowels removed → `l.ngt.mps`
- withdrawal of [n, m] before consonant → `l..gt..ps`
- withdrawal of [l, r, h] after consonant → `l..gt..ps`
- ⇒ **lgtps**

Transformation of a word to a Consonant Skeleton (2)

- Some word can not be shortened this way
  - `oiseau` → **oiseau** (*bird*)
- Silent letters might be kept
  - `longtemps` → LGTPS (*long* [for a long time])
  - `toujours` → TJRS (*always*)
- Not specific to SMS
  - Exists in several languages
  - Some occurrence are stable (`dvlpt` – développement, development)

Transformation of a word to a Consonant Skeleton (2)

- Some word can not be shortened this way
  - `oiseau` → **oiseau** (*bird*)
- Silent letters might be kept
  - `longtemps` → LGTPS (*long* [for a long time])
  - `toujours` → TJRS (*always*)
- Not specific to SMS
  - Exists in several languages
  - Some occurrence are stable (`dvlpt` – développement, development)

Transformation of a word to a Consonant Skeleton (2)

- Some word can not be shortened this way
  - `oiseau` → **oiseau** (*bird*)
- Silent letters might be kept
  - `longtemps` → LGTPS (*long* [for a long time])
  - `toujours` → TJRS (*always*)
- Not specific to SMS
  - Exists in several languages
  - Some occurrence are stable (`dvlpt` – développement, development)

Lexicon of Consonant Skeleton

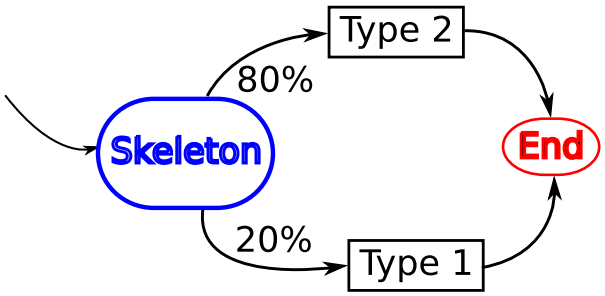
Building a lexicon of Consonant Skeleton:

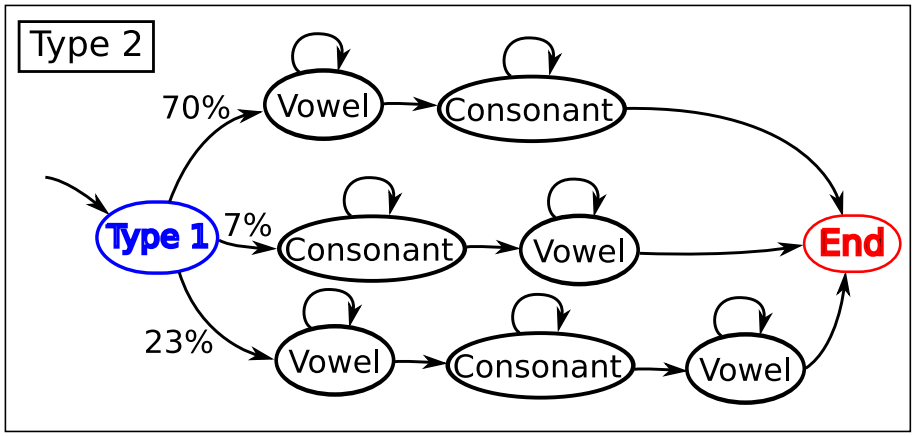
1. Starting from *Le Monde* newspaper corpora
2. Selecting nouns, adverbs and adjectives (frequency above choosen threshold) $\longrightarrow$ 3244 word processed
3. Applying transformation rules to this selection
4. $\Rightarrow$ list of Consonant Skeleton

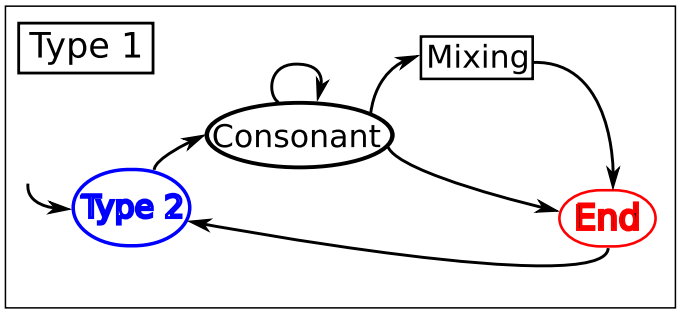
Regular Expression for Consonant Skeleton

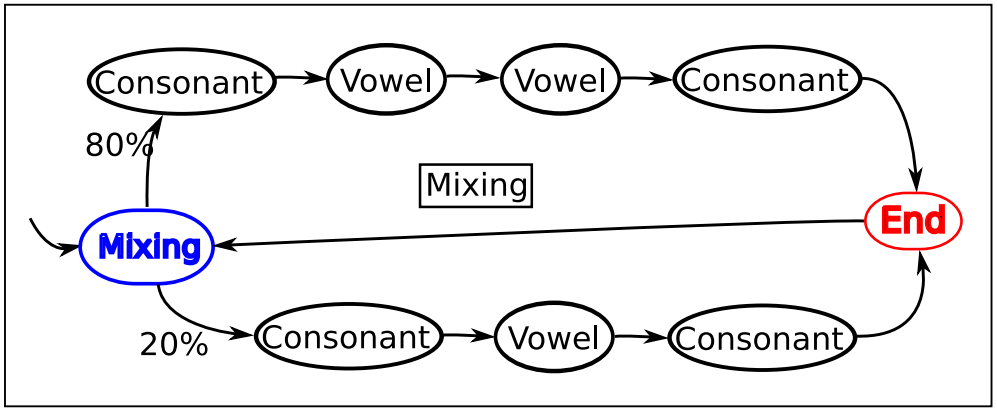
Characterize the shape of a consonant Skeleton

- words are mostly composed of vowels ;
- some exceptions (beginning and end of word) ;
- some consonant removed anyway.
- possibility to keep some vowels for partially shortened word (ex `bjour` – *bonjour, hello*)


Type 2
80%
Skeleton
End
20%
Type 1


Type 2
70%
Vowel
Consonant
Type 1
7%
Consonant
Vowel
End
23%
Vowel
Consonant
Vowel


Type 1
Mixing
Consonant
Type 2
End


Consonant
Vowel
Vowel
Consonant
80%
Mixing
End
Mixing
20%
Consonant
Vowel
Consonant

Results for Consonant Skeleton

Lower Bound (char)	94,7%
Lower Bound (word)	85,2%
RegExp (char)	98,0%
RegExp (word)	94,4%
Lexicon (char)	94,7%
Lexicon (word)	85,2%
RegExp+Lexicon (char)	98,0%
RegExp+Lexicon (word)	94,4%
Upper Bound (char)	100%
Upper Bound (word)	100%

Results for Consonant Skeleton

Lower Bound (char)	94,7%
Lower Bound (word)	85,2%
RegExp (char)	98,0%
RegExp (word)	94,4%
Lexicon (char)	94,7%
Lexicon (word)	85,2%
RegExp+Lexicon (char)	98,0%
RegExp+Lexicon (word)	94,4%
Upper Bound (char)	100%
Upper Bound (word)	100%

Results for Consonant Skeleton

Lower Bound (char)	94,7%
Lower Bound (word)	85,2%
RegExp (char)	98,0%
RegExp (word)	94,4%
Lexicon (char)	94,7%
Lexicon (word)	85,2%
RegExp+Lexicon (char)	98,0%
RegExp+Lexicon (word)	94,4%
Upper Bound (char)	100%
Upper Bound (word)	100%

Other phenomena

- Rebus processed using Regular Expression
  - No improvement
- Phonetic writing processed using Lexicon
  - Slight improvement at word level

Conclusions

- Limited resources available
  - Results to be confirmed (see `http://www.smspourlascience.be/`)
- A first step toward SMS characterization
  - Improve and Validate
- Next move: processing combination of phenomena
  - Recognition rate slightly increased for isolated phenomena
  - $\rightarrow$ Not sufficient to process complex forms
  - Example: `2nite` – *tonight*, combination of Rebus and Phonetic Writing

Conclusions

- Limited resources available
  - Results to be confirmed (see `http://www.smspourlascience.be/`)
- A first step toward SMS characterization
  - Improve and Validate
- Next move: processing combination of phenomena
  - Recognition rate slightly increased for isolated phenomena
  - → Not sufficient to process complex forms
  - Example: `2nite` – *tonight*, combination of Rebus **and** Phonetic Writing

Conclusions

- Limited resources available
  - Results to be confirmed (see `http://www.smspourlascience.be/`)
- A first step toward SMS characterization
  - Improve and Validate
- Next move: processing combination of phenomena
  - Recognition rate slightly increased for isolated phenomena
  - $\longrightarrow$ Not sufficient to process complex forms
  - Example: `2nite` – *tonight*, combination of Rebus **and** Phonetic Writing

Thanks !

Recognition Rate

- D : Levenshtein distance compute between original text and recognised text
- Insertion cost $= 0$
- Deletion/substitution cost $= 1$
- Example:

Label:	`bjr`	(taille: 3)
Recognized:	`loj.t`	(taille: 5)
Distance:	2	
Precision:	$3 - 2 = 1 \rightarrow 1/3 = 33\%$	

$$\Rightarrow RR = 100 \times (\#label - D)/\#label)$$