Stentor

A new Computer-Aided Transcription software for French language

SténoMédia

- Young compagny : 6 months
- Pretty long history : 6 years and 2 years
- Association of professional and researchers
  - S. Badot : International stenotypist
  - Y. Estève : Researcher, Université du Mans
  - T. Spriet : Researcher, Université d'Avignon
- Stenotypy domain : a very good lab

SténoMédia

French Stenotypy

- "Grandjean" method, based on words pronunciation
- High rate of homophony, about 1.8
- long span syntactic constraints
- various application domains

SténoMédia

Stenotypy vs Speech recognition

- ***"acoustic variations" due to typing errors***
- ***"acoustic similarities" due to ambiguities of the French stenotypy method***
- ***high rate of homophones, increased by ambiguities provided by the stenotypy method***
- ***high rate of homographs, which cannot be efficiently reduced by a n-gram model***

SténoMédia

Stenotypy vs Speech recognition

- human interpretation
  - ***delete stammering and hesitations***
  - ***speaker identification***
  - ***add some extra speech events***
- ***punctuations, breakpoints***

specificities

- Very large vocabulary and specific lexicons
- New words in realtime
- Personnal adaptations
- independence of the stenotypy method

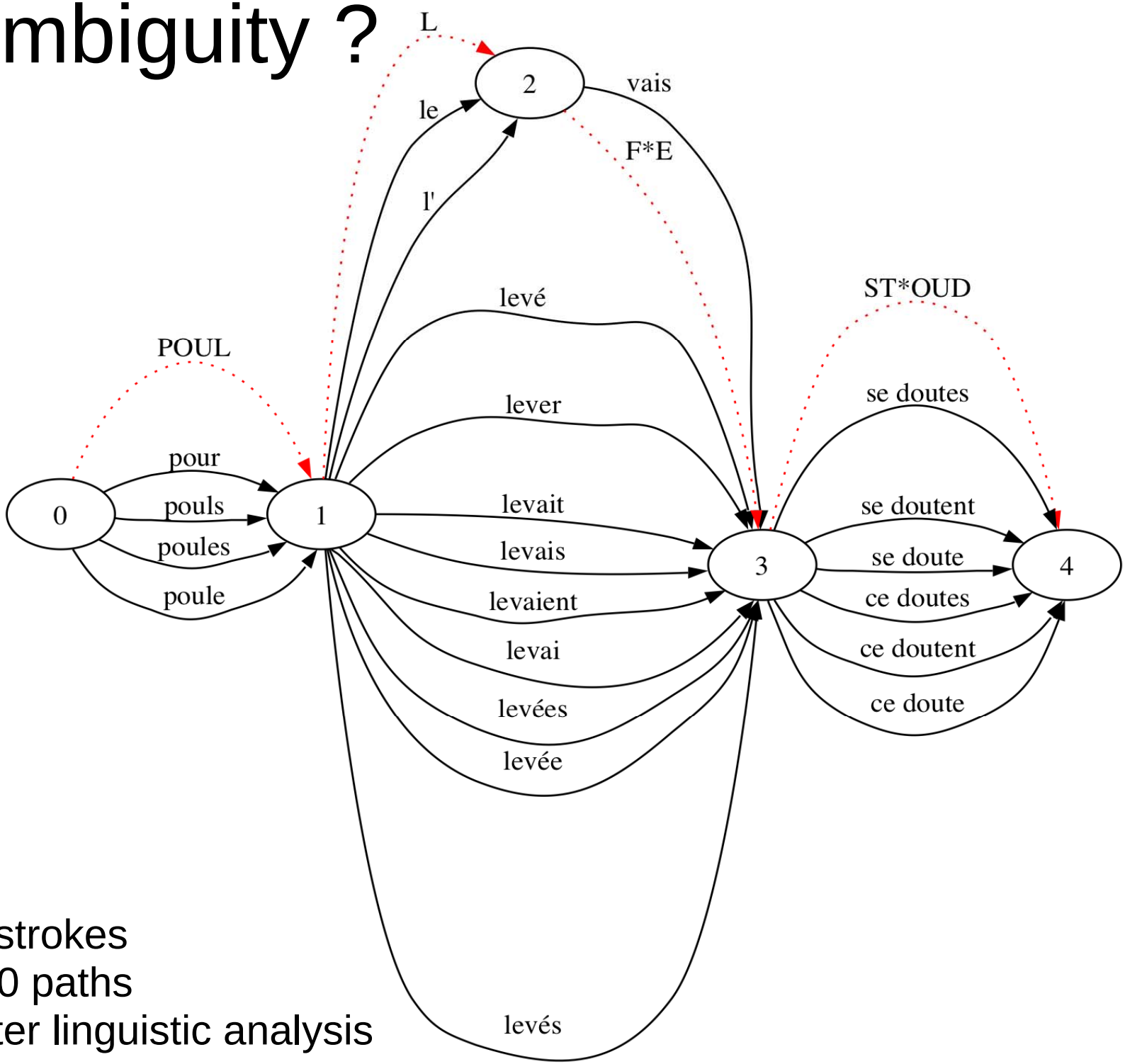
SténoMédia

Personnal adaptations

- New stenotypy of a word
- Syntactic class modification
- Words concatenation
- post treatments

SténoMédia

Ambiguity ?



example 4 keystrokes
- more than 360 paths
- only 4 or 5 after linguistic analysis

SténoMédia

Linguistic model

- Mixte approach :
- Linear combination of language models
  - 3-gram
  - 3-class
- knowledge rules

3 gram

- The 3-gram model is in fact a combination of 3-gram, 2-gram en 1-gram models
- Trainning corpus about 4.5 millions of words
- Lexicon about 150K most used words

3 class

- statistics association of Part Of Speech (POS)
- tag set of 105 POS
  - NMS,
  - VA1PS,
  - XSOC,
  - ...

Knowledge rules

- Used for special forms like
  - words after a '*apostrophe*' must begin with by a vowel
  - The first word of the sentence must have capital
  - a verb after '*pour*' is in infinitive form

Results

- NIST Scoring Toolkit (SCTK)
- only a 5K words in test corpus manually corrected
- word error rate of 5% (1% earned this month)

Conclusion

- The first version of STENTOR is now out and is used by the profession
- World error rate already competitive
- Some improvements planned
  - long span dependencies
  - a better dictionnary
  - a larger training corpus

Conclusion

- Stentor a good lab but also a professional software with
  - Audio-sync
  - dictionaries builder
  - realtime word insertion
  - computer assisted correction
  - short cuts

Sténo Média

Questions ?

Thank you for your attention