Christophe Van Bael

3rd Euromasters Summerschool

Project Presentation

Using the keyword lexicon for speech recognition

8 July 2002

This talk

- Architecture of a speech recogniser
- Past usage of lexica for speech recognition
- The keyword lexicon — a new approach
- Experiments: Done's and To Do's
- Conclusion

Architecture of a speech recogniser

Basic view:

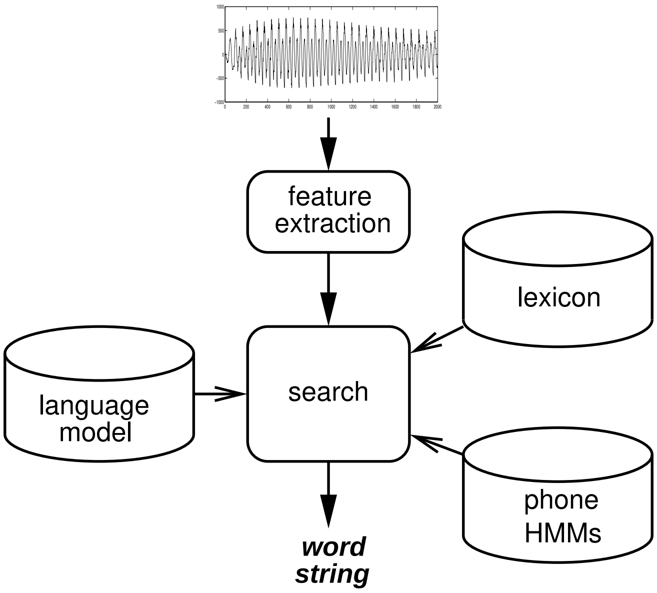
- speech recognition as decoding
- training and recognising
- constraints on search space

These constraints include:

- language model
- word model
- phone model

Architecture of a speech recogniser

Block diagram:



(King, 2001)

Models and their contribution to accuracy:

- the models shouldn't be too strict
- but tight enough to narrow down the search space

Past usage of lexica for recognition

Typical characteristics of lexica:

- multiple pronunciations per lexical entry

BATTERER	b ae t ax ax
BATTERER	b ae t ax ax r
BATTERER	b ae t ax r ax
BATTERER	b ae t ax r ax r

(example entries taken from BEEP BrE lexicon)

- examples:
  - CMU, AmE lexicon
  - BEEP, BrE lexicon

Past usage of lexica for recognition

Why do we need/want alternative pronunciations:

- different speakers have different pronunciations
- this pronunciation variance can be covered
- and if enough data available this means more specific models can be trained

Why then bother with new approaches?

Past usage of lexica for recognition

Disadvantages of alternative pronunciations:

- if not enough data poor models will be trained
- adding more alternative pronunciations improves accuracy. . . up to a certain point!
- after that point, accuracy will go down again because
  - bigger chance for homophones in lexica
  - more confusion (bad recognition) possible
  - *weak constraint on search space!*

Isn't there a clever way around this?

The Keyword Lexicon

Exploit the fact that speakers are consistent in their speech (at least most of the times)

- if a speaker starts of with *b aa th* (Bath, placename UK), chance is low for a switch to *b ae th* .

- and within one speaker, generalisations can be made

- if a speaker starts of with *b aa th* , then *p aa th* will be highly likely

- generalisation between groups of speakers (different regional accents) possible

Standard recognisers haven't exploited these consistencies. . .

until now!

The Keyword Lexicon

Wells (1982) and the concept of *standard lexical sets* or *keywords*

- because matching pronunciations of words in different accents is often difficult

- ease of reference

list of original keywords (originally only keyvowels)

- kit, dress, trap, lot, strut, foot, bath, cloth, nurse, fleece, face, palm, thought, goat, goose, price, choice, mouth, near, square, start, north, force, cure + happY, lettER, commA

The Keyword Lexicon

The key idea:

- John C. Wells (Accents of English, 1982): the keyvowel idea

The implementation for speech synthesis:

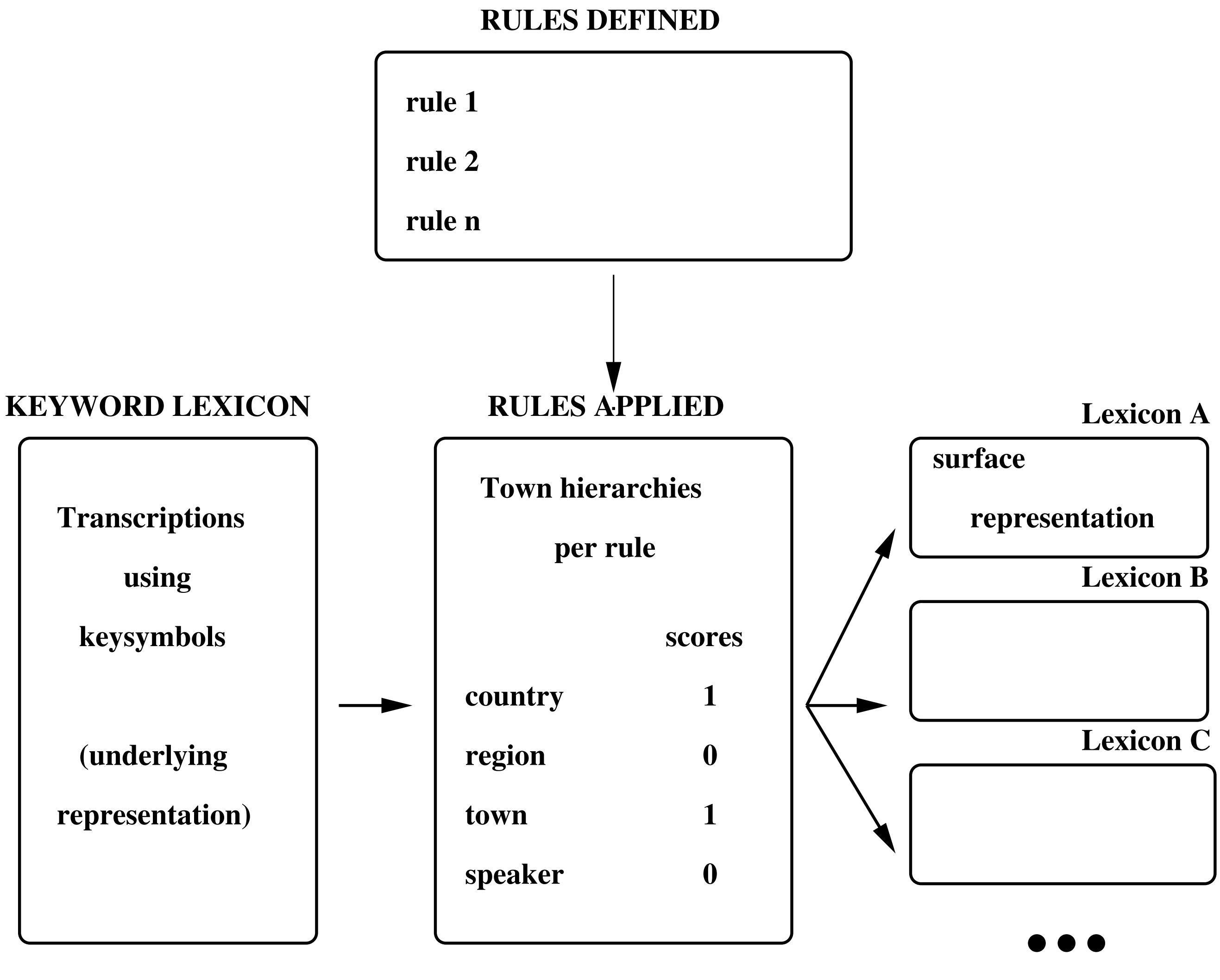
- Sue Fitt at CSTR: Keyword Lexicon built as a new way to model pronunciation variation

The implementation for speech recognition:

- Christophe Van Bael and Simon King: Keyword Lexicon as a constraint on recognition

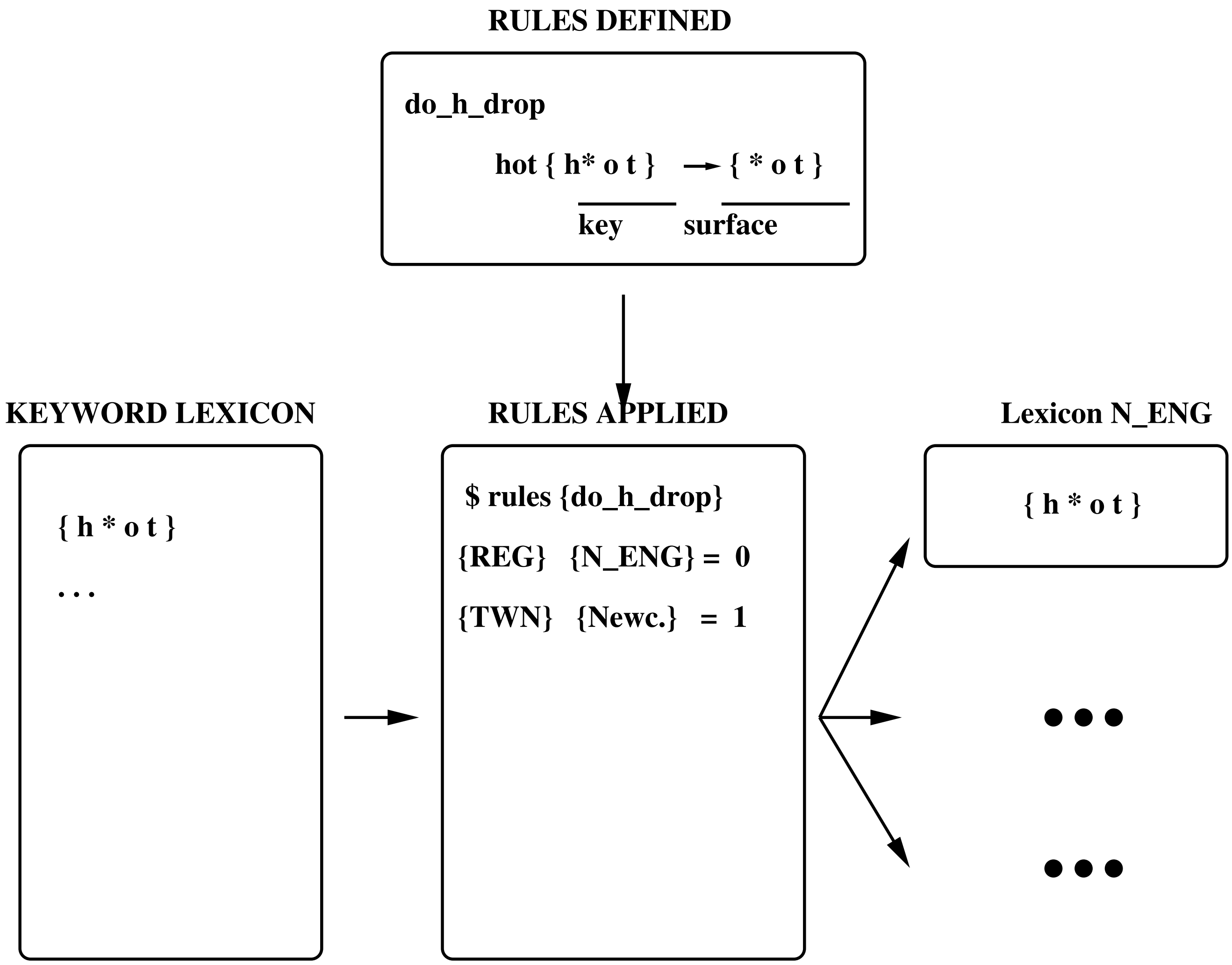
The Keyword Lexicon

Basic architecture:



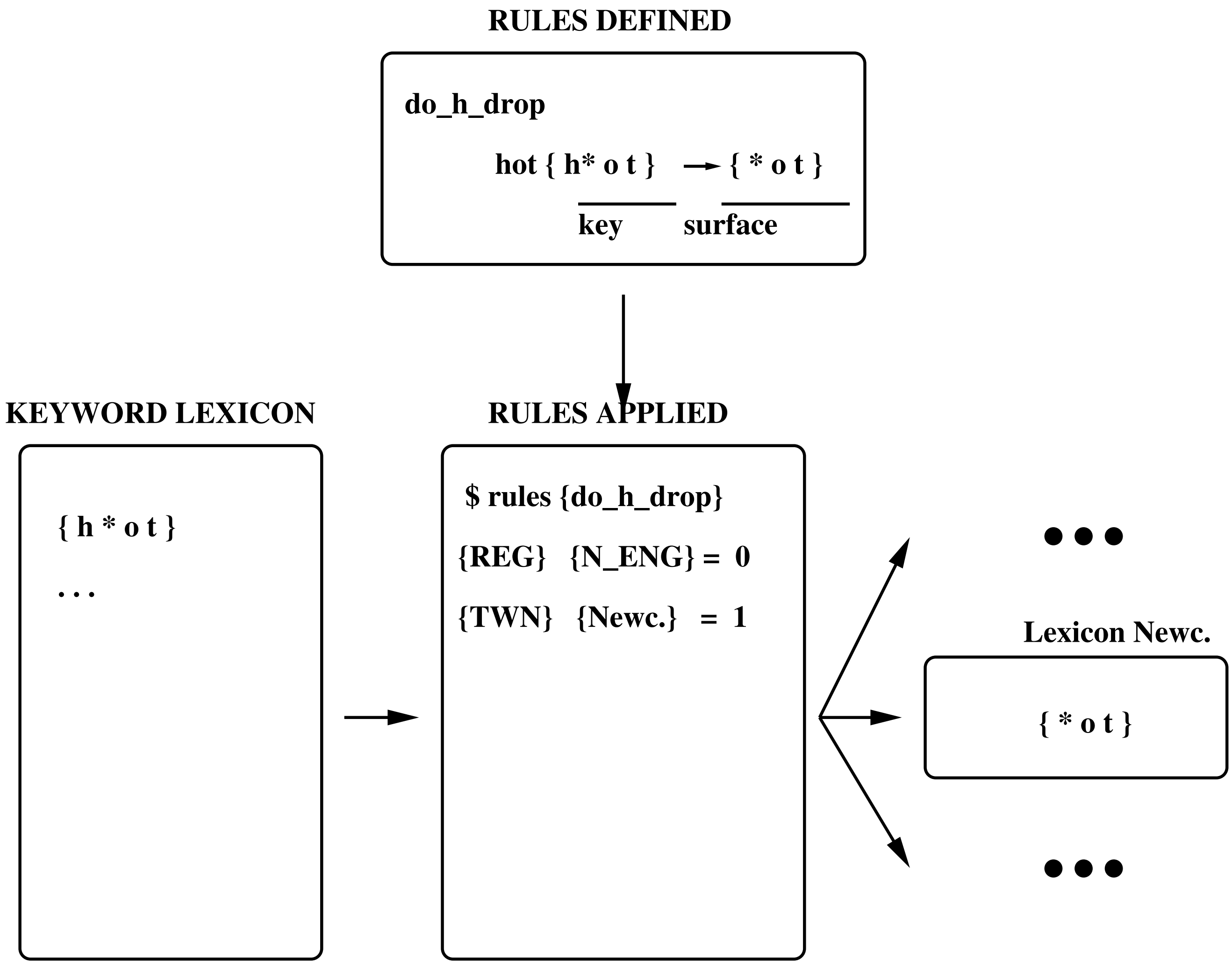
The Keyword Lexicon

An example (1/3):



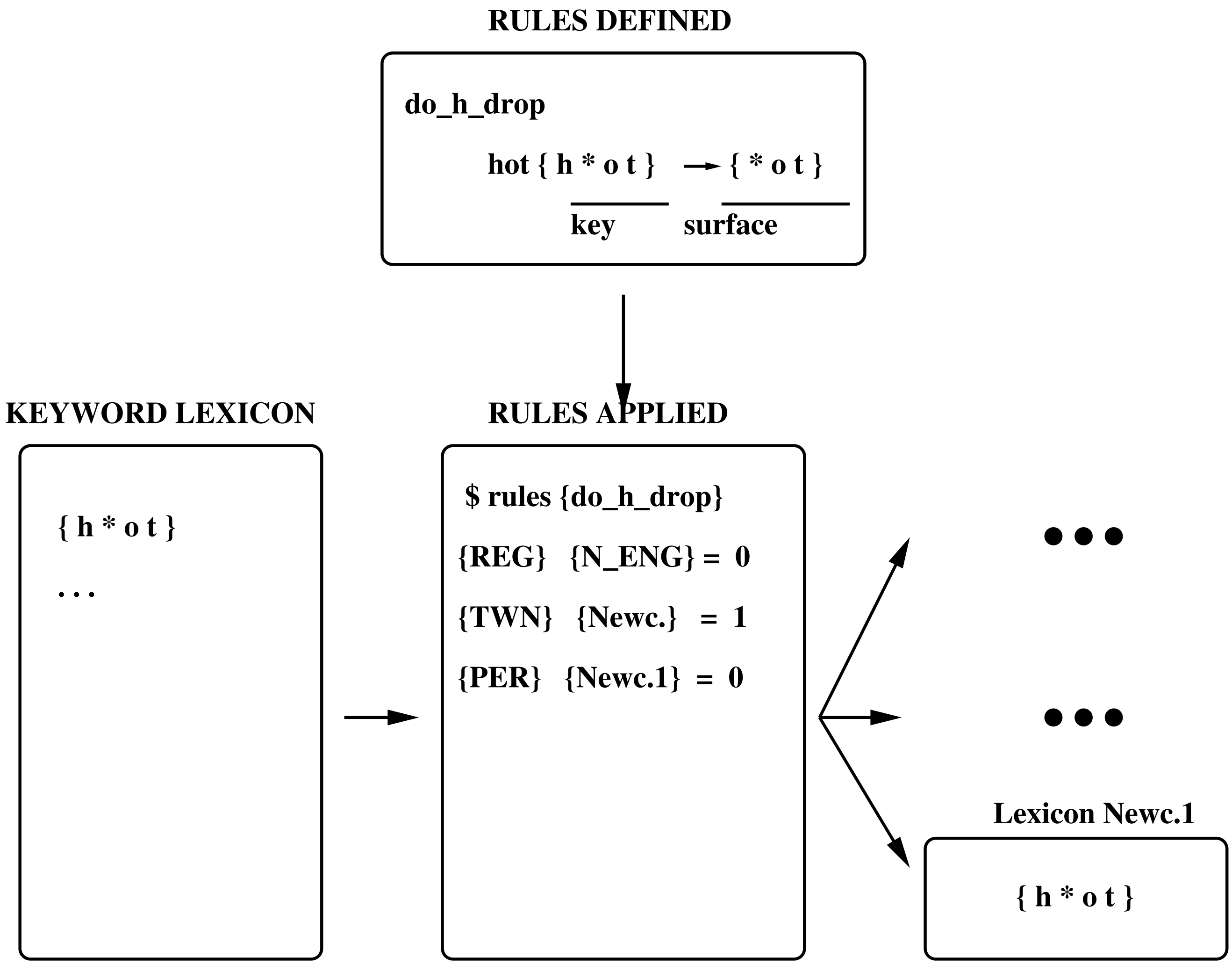
The Keyword Lexicon

An example (2/3):



The Keyword Lexicon

An example (3/3):



The Keyword Lexicon

Keyword lexicon for recognition:

- reason 1: less confusability/better constraint on search space
- reason 2: reduces memory requirements

Experiments:

- baseline: standard in literature: standard multi-pronunciation lexicon on a standard task (BEEP on WSJCAM0): 11,4% WER (Robinson et al., Cambridge University)
- experiment 1: Approach that baseline same lexicon on the same standard task: 34,08% WER (different tools)
- experiment 2: With same parameters merged Keyword Lexicon on the same standard task:
- experiment 3: With same parameters real Keyword Lexicon on the same standard task:

Done's and To Do's

Done so far:

- trying to approach standard WER from literature
  - BEEP on WSJCAM0, 5K task
  - triphones, 4 gaussian mixtures, within word models, standard bigram LM, average pruning
  - accuracy on small 200 sentences development test set: 34,08% WER
- trying new approach by using Keyword Lexicon
  - merged (so far) Keyword Lexicon on WSJCAM0, 5K task
  - triphones, 15 gaussian mixtures, within word models, standard bigram LM, average pruning
  - accuracy on small 200 sentences development test set:

Done's and To Do's

Still to do:

- get the real Keyword Lexicon + rules approach working
- compare old and new approach in terms of WER
- test everything on big evaluation test set

Conclusion

speech recognition: narrowing down the search space

use of lexica until now: multiple alternative pronunciations

- major advantage: capture pronunciation variation

- disadvantages: confusability, weak constraint on search space + effort and storage space

better to exploit similarities or consistencies in pronunciation

- new approach: Keyword Lexicon (first for synthesis, now also for recognition)

Conclusion

experiments:

- experiment 1 : BEEP baseline on small development test set: 34,08% WER
- experiment 2 : merged Keyword Lexicon baseline on small development test set:

still to do:

- experiment 1 : BEEP baseline on evaluation test set
- experiment 2 : merged Keyword Lexicon baseline on evaluation test set
- Keyword Lexicon + rules instead of merged lexicon