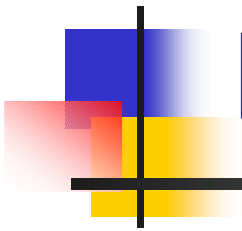
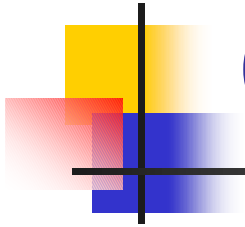


FipsOrtho: A Spell Checker for Learners of French

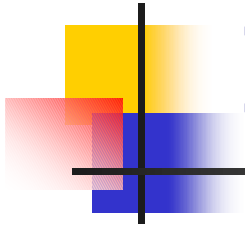
Sébastien L'haire

University of Geneva
Switzerland



Outline

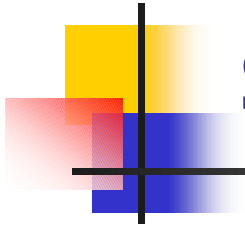
- Issues of FipsOrtho
- Techniques involved
- Demo
- Results of testing
- Future plans



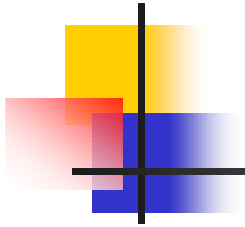
Issues of FipsOrtho

- Specific mistakes done by learners
- Web application PHP + MySQL + CGI program
- Specific techniques:
  - Syntactic analysis (at least chunks)
  - Phonetisation
  - Ad hoc rules
  - Ordering of proposals
  - XML output parsed by PHP
- Gathering corpus of authentic xml-tagged learners productions

Spell checking: les travails sont difficiles.

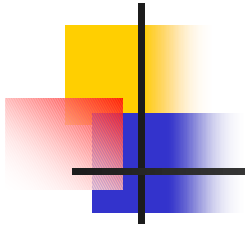


- Analysis [S [NP Det Les N* travails] [VP V sont [AP difficiles]]] → unknown word travails* should be a plural noun!
- Alphacode method:
  - string of all consonants by alphabetical order followed by vowels: lrstvai
  - Narrowing: rstvai, lstvai etc.
  - Widening: blrstvai, clrstvai etc.
  - 27 queries to lexicon
- Results: travail (N), travailla (N), travaillai (N), travaillais, travaillas, travaillasse (W), travaillât (N), travaillées (W), travaillés (W), travailles (W), allitératives (W), ravitaillais etc. A=6, W=93, N=49, Total=148



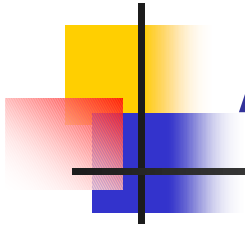
Filtering

- Only proposals beginning by same character
- Measuring of lexicographic distance (number of insertions / deletions / inversions / substitution necessary to go from string A to B) → threshold to eliminate too different strings
- Adaptation of Levenshtein / Damerau distance
  - Confusion between single / double consonant less penalised
  - Confusion of diacritics less penalised
- Retains: A=2, W=4, N=4, Total: 10 proposals



Phonetisation

- Unknown word phonetised
- Retrieves travail, travaille, travaillent, travailles
- 2 new proposals



Ad hoc rules

- Endings with –ails and -als replaced with –aux, -age with –ment and –ment with –age (substantives)
- Beginings with aller- replaced by ir-, tenir- with tiendr-, voir- with verr-, fair- with fer- (verbal futures)
- Retrieves travaux

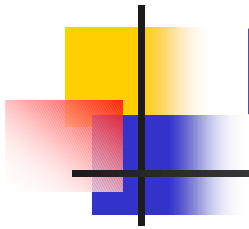
Apostrophe, separation, case

- Qu’, c, d, j, l, m, n, s and t can be followed by apostrophe
- Existing words w/ apostrophe: aujourd’hui, prud’homme, prud’hommal, presqu’île
- Try to separate words by inserting space
- First word in capital letter

Ordering

- Score by method(s) involved + fit into parser's analysis (person, gender, number)
- travaux is a plural noun → fits best to the analysis
- Global result: 13 proposals

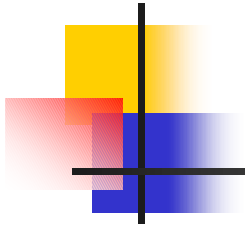
1. travaux
2. travail
3. travailles
4. travaillés
5. travaille
6. travaillas
7. travaillent
8. travailla
9. travaillées
10. travaillais
11. travaillasses
12. travaillai
13. travaillât



DEMO

First experiment:154 words

Methods	Score	Percent
Alpha + phono	44	28.57
Alpha widened + phono	12	7.79
Alpha narrowed + phono	7	4.55
Alpha	33	21.43
Alpha widened	6	3.9
Alpha narrowed	6	3.9
Phono	23	14.94

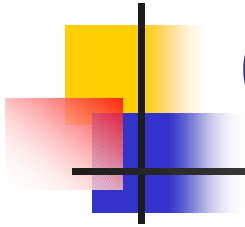


First experiment (cont.)

- 12 words with no proposals (7.79%)
- 11 words without correct proposal (7.14%)
- Average: 146.4 words retrieved but 5.955 words selected by filtering
- Alpha yields average 11.7 results, narrowed alpha 54.7 and widened alpha 79.4

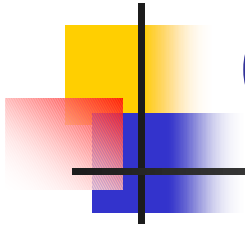
Corpus result

- 295 entries from
  - Authentic sentences provided by teachers (Jamaïca, Australia and Canada)
  - Articles on CALL
  - Sentences from « Dictée de Bernard Pivot » filled with errors (benchmark of grammar checkers)
  - Email from a native speaker



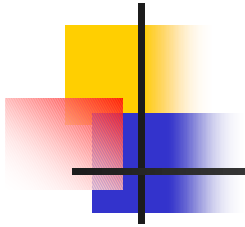
Corpus result (2)

- 6651 words, 558 sentences, average 22.55 w/entry
- 1107 errors tagged (avg. 3.75 /entry)
- 404 unknown words.
  - 196 corrected by techniques
  - 159 left unchanged (proper names, unknown words, false detections), 87 of them had proposal(s)
  - 49 detected errors corrected by hand (unknown words, 32 had proposal(s)



Corpus result (3)

- 703 errors undetected
  - Agreement errors (171)
  - Complementation (89)
  - Superfluous word (64)
  - Missing word (59)
  - Etc.
- Necessity of combining spell with grammar and stylistic checking



Future plans

- Testing with learners
- Deep analysis of errors → improvement
- Improvement of results ordering
- Improvement of phonetisation (adaptation to L1??)
- Improvement of ad-hoc method → true morphological analysis
- Learning tools
  - Access to lexicon with subcategorisation frames
  - Morphological analysis
  - Phonetic transcription of words and speech synthesis
  - Various interfaces adapted to level?