INTRODUCTION

■ Abstract
“This paper presents a spell checker catering to the specific needs of language learners. Spell checking consists in verifying that each word of a text belongs to the language in which the text is written and, if need be, in proposing possible corrections.”

■ What is FipsCor?
  a) FipsCor is a French spell checker built specifically for language learners.
  b) The system uses a dictionary of approximately 200,000 word forms.
  c) Each word form contains its alpha-code, its phonological representation, and its lexical category among others.
  d) It stores complete word forms, rather than roots and inflection rules.
  e) Thus, a morphological analyzer can not be used to access the dictionary.

■ Why is a spell checker needed especially for language learners?
  a) Type of errors that language learners make is different from that of native speakers.
     i) Native speaker errors: mainly typing errors (up to 80%).
        Ex) letter transposition, insertion, deletion, and substitution.
     ii) Language learners: other types of errors added to those of native speakers.
        Ex) errors from relying too much on the sounds of the language.
  b) Specific features needed for language learners
  c) Commercial spell checkers are rarely specific for language learners.
  d) By privileging ad hoc (morphology) and phonological reinterpretation methods, it promotes correction proposals for spelling errors specific to language learners.

■ Three methods for retrieving correction proposals are discussed:
  1) Alpha-Code Method (a well-known technique)
  2) Phonological Reinterpretation Method (specific for language learners)
  3) Ad Hoc Method (specific for language learners)

THE THREE METHODS

1. **Alpha-Code Method**
   a) Definition – an alpha-code is an abstract representation of a word formed by the characters which compose it.
   b) It allows the system to provide a correction for errors of reduplication, inversion, some occurrences of letter deletion and insertion, as well as missing diacritics.
   c) It is a well-known technique adequate for many typing errors and often used in spell checkers.
   d) The version of the alpha-code used in FipsCor:
- The alpha-code is formed by the consonants of the word ordered alphabetically, followed by the vowels, also in alphabetical order. Each letter is kept only once.
  
ex) table → bltae prefer → fpred experience → cnprxe
  
e) Many words can share the same alpha-code, however.
  
ex) able → blae label → blae bale → blae

**2. Phonological Reinterpretation**

a) Language learners tend to make types of errors other than the typing errors native speakers make.

b) Learners might try to write some words phonetically and use an incorrect grapheme-to-phoneme correspondence. This is called ‘phonetic writing’.

c) Phonological reinterpretation method is specifically for learners.

d) It is built upon the source of the learner error in order to retrieve correction proposals which would otherwise have been lost.

  e) i) The system first tries to discover the pronunciation of the erroneous word.

  ii) It uses a phonetiser instead of a dictionary to select the phonological representation of the word because ‘phonetic writing’ is not part of the dictionary.

  iii) The phonological representation is used as the search key to retrieve all the words with the same pronunciation of the erroneous word.

  iv) The retrieved words serve as correction proposals.

**3. Ad Hoc Method**

a) The spell checker uses an ad hoc technique for a highly frequent, but quite specific, error type: the incorrect plural formation of words

  Ex) in French *animals instead of animaux

  Ex) in English *childs instead of children *gooses instead of geese

b) A morphological analyzer cannot be used because the dictionary stores complete word forms, rather than roots and inflection rules.

c) The ad hoc method permits the bypassing of a morphological analysis.

d) Other rules of this kind for other types of errors could be designed in order to propose possible corrections for erroneous words due to morphology.

<table>
<thead>
<tr>
<th>Method</th>
<th>Weakness</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha-code</strong></td>
<td>Too many correction proposals due to the fact that many words can share the same alpha-code</td>
<td>Use lexicographic distance threshold</td>
</tr>
<tr>
<td><strong>Phonological reinterpretation</strong></td>
<td>Only retrieves words with the exact same pronunciation &amp; pronunciation errors dependent on learners’ L1</td>
<td>Relax method to include near homophones &amp; parameterize relaxation to L1</td>
</tr>
</tbody>
</table>

**Weaknesses of the three methods**
<table>
<thead>
<tr>
<th>Ad hoc</th>
<th>Impossible to cover all the types of errors</th>
<th>The use of a morphological analyzer</th>
</tr>
</thead>
</table>

- **Testing the FipsCor**
  - a) **FipsCor**’s current performance was compared to the spell checker included in Microsoft Word, version 2000.

Table 3. p.224

<table>
<thead>
<tr>
<th></th>
<th><strong>FipsCor</strong></th>
<th></th>
<th><strong>Word 2000</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>%</td>
<td>Number</td>
<td>%</td>
</tr>
<tr>
<td>Detected</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct proposal</td>
<td>112</td>
<td>77.78</td>
<td>136</td>
<td>94.44</td>
</tr>
<tr>
<td>Incorrect proposal</td>
<td>13</td>
<td>9.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No proposal</td>
<td>18</td>
<td>12.5</td>
<td>1</td>
<td>0.69</td>
</tr>
<tr>
<td>Undetected</td>
<td>1</td>
<td>0.69</td>
<td>7</td>
<td>4.86</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>100</td>
<td>144</td>
<td>100</td>
</tr>
</tbody>
</table>

- b) Table 3 analysis
  - i) At first sight, **FipsCor** seems to do worse than Word 2000 because it proposes the proper correction only 77.78% of the time, while Word 2000 nearly reaches 95%.
  - ii) The number of ‘incorrect proposal’ and ‘no proposal’ is higher for **FipsCor** reinforcing the first impression.
  - iii) However, it is crucial to note:
    - The number of ‘undetected words’ for Word 2000 is higher than that of **FipsCor**.
    - Of the 31 instances of ‘incorrect proposal’ and ‘no proposal’, 27 of them were due to the fact that the correct words were missing from the **FipsCor** dictionary.
    - With the addition of those 27 words, **FipsCor** would have a higher percentage of approximately 96%, contrary to the first impression.
    - Although the correct form is included in the dictionary, the four remaining words were ‘incorrect proposal’ or ‘no proposal’ due to a combination of several reasons:
      1. Letter substitution error
      2. Lexicographic distance threshold too low to include correction proposal.
      3. Phonetiser did not propose a phonological representation of the erroneous word close enough from the pronunciation of its correction.

**CONCLUSION**

- **Further developments needed**
  - a) Update dictionary
    - Enlarge the dictionary
    - Correct any errors
    - Add phonetiser

b) Develop a real morphological treatment instead of the ad hoc method
   - Ad hoc method too specific
   - A morphological analysis would ensure a much larger coverage of errors learners produce.

c) Vary lexicographic distance threshold depending on the number of alternatives retrieved.
   - Some correct proposals could be eliminated due to too low threshold.
   - Raise the threshold if there are too few correction proposals.

d) Order the correction proposals
   - Shortlist the proposals and/or order them according to:
     1) Proposals found by the ad hoc (or later a more complete morphological treatment) are displayed first.
     2) Next, proposals found by phonological reinterpretation, then those found by the alpha-code method.
     3) Proposals found by two methods should be promoted to the top of the list.
     4) Within the list of proposals from the alpha-code and the phonological reinterpretation, ordering is based on the lexicographic distance.

e) Link FipsCor to a syntactic parser
   - Parse the sentence and filter the list of correction proposals to allow only those that fit in the sentence syntactically.

Further thoughts...

a) It is important to use several methods to find correction proposals and to combine them in order to be able to retrieve the maximum number of possible corrections, as it is the only way to make as certain as possible to include in our list the valid correction needed by the user.
b) It needs to be tested more thoroughly on a large sample of authentic learner data, in order to attest that it is well adapted to its target users, language learners.

Definitions:

a) Lexicographic distance: It computes “the minimum number of editing operations (insertion, deletion, and substitution) needed to transform one string into another.” It acts only as a filter on the words pre-selected by the alpha-code method.
   - Ex) letter insertion and deletion is assigned a cost of 1
   - letter substitution is assigned a cost of 2
b) Phonetiser: It is “a classic rule-based system (about 700 rules), in which the rules are selected according to the graphemic sequence, graphemic left and right contexts, as well as syntactic context.”
c) To parse: to analyze a sentence using a grammar, including deciding whether it is valid and what its structure is according to the grammar.