Automated Writing Assistance: Grammar Checking and Beyond
Topic 1: The Nature of the Problem

Robert Dale
Centre for Language Technology
Macquarie University
Motivating Questions

• What does it mean to help people write?
  – What kind of help would you like with your writing?

• How would we know if we had succeeded in helping someone write?
  – What does it mean to improve a piece of writing?
  – What does it mean for a piece of writing to be good?
What This Course is About

• How we can use NLP tools and techniques to help people write:
  – Spell checking
  – Grammar checking
  – Style checking
  – Discourse-level assistance
What This Course is Not About

- Teaching or helping with handwriting
- Teaching how to type
- Teaching a language
- Productivity tools like editors and word processors
Overview

• The Nature of the Writing Process
• Categorising Errors
The Conduit Metaphor #1
The Conduit Metaphor #2

"Thelma and Louise"

"let's go see a movie"

"Terminator"
Kinds of Writing

- Fiction
- Poetry
- Personal journals or narratives
- News reporting
- Technical writing
- ...

SSLST 2011
... the primary, though certainly not the sole, characteristic of technical and scientific writing lies in the effort of the author to convey one meaning and only one meaning in what he says. That one meaning must be sharp, clear, precise. And the reader must be given no choice of meanings; he must not be allowed to interpret a passage in any way but that intended by the writer.

Britton 1965:114
A Stage Model of the Writing Process

Rohman 1965
A Cognitive Process Model

Flower and Hayes 1981
The Nature of Revision

Figure 1. A Taxonomy of Revision Changes

Faigley and Witte 1981
Additions make explicit what can be inferred:

- you pay two dollars → you pay a two dollar entrance fee
Meaning-Preserving Changes: Deletions

Deletions remove explicit elements and force the reader to infer:

• several rustic looking restaurants → several rustic restaurants
Substitutions replace words or phrases with other synonymous content:

- out-of-the-way spots → out-of-the-way places
Meaning-Preserving Changes: Permutations

Permutations rearrange material, possibly with substitutions:

• springtime means to most people
  → springtime, to most people, means
Distributions move material from one segment into multiple segments:

- I figured after walking so far the least it could do would be to provide a relaxing dinner since I was hungry.

→

I figured the least it owed me was a good meal. All that walking made me hungry.
Consolidations move material from multiple units into a single unit:

- And there you find Hamilton's Pool. It has cool green water surrounded by 50-foot cliffs and lush vegetation.

→

And there you find Hamilton's Pool: cool green water surrounded by 50-foot cliffs and lush vegetation.
The Nature of Revision

Revision Changes

Surface Changes

- Formal Changes
  - Spelling
  - Tense, Number, and Modality
  - Abbreviation
  - Punctuation
  - Format

- Meaning-Preserving Changes
  - Additions
  - Deletions
  - Substitutions
  - Permutations
  - Distributions
  - Consolidations

Text-Base Changes

- Microstructure Changes
  - Additions
  - Deletions
  - Substitutions
  - Permutations
  - Distributions
  - Consolidations

- Macrostructure Changes
  - Additions
  - Deletions
  - Substitutions
  - Permutations
  - Distributions
  - Consolidations

Figure 1. A Taxonomy of Revision Changes

Faigley and Witte 1981
Meaning Changes

• **Macrostructure changes**
  – Would change a summary of the text
  – Impact on reading of other parts of the text

• **Microstructure changes**
  – Don’t change the gist of the text
  – Are isolated in impact
Writing Assistance: The State of the Art

- Existing tools are concerned with surface revisions, and even then primarily with formal changes
- We can conceive of machine assistance being provided for every aspect of revision
- We can also conceive of machine assistance being provided for the prewriting and writing stages
Writing Assistance in the Future?
The Clippy Problem

- Microsoft’s Office Assistant
- Technical advances will need to address workflow integration issues
Overview

- The Nature of the Writing Process
- Categorising Errors
Common Errors in English
Ways of Categorising Errors

• By cause:
  – Mechanical errors [also known as errors of execution or performance errors]
  – Cognitive errors [also known as errors of intention or competence errors]

• By symptom:
  – Misspelled words
  – Ungrammatical sentences
  – Stylistic disfluencies and inconsistencies
An Analysis of Student Writing Errors

• Connors and Lundsford 1988:
  – 21,500 corrected student papers from 300 teachers across the USA
  – 30% typed, 70% handwritten
  – Length varied from less than a page to over 20 pages
  – Randomly selected 3000 for analysis
A Taxonomy of Errors

• Developed on the basis of an analysis of 300 papers

<table>
<thead>
<tr>
<th>Error or Error Pattern</th>
<th>#</th>
<th>Error or Error Pattern</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td>450</td>
<td>Subject-verb agreement</td>
<td>59</td>
</tr>
<tr>
<td>No comma after introductory element</td>
<td>138</td>
<td>Unnecessary comma with restrictive phrase</td>
<td>50</td>
</tr>
<tr>
<td>Comma splice</td>
<td>124</td>
<td>Unnecessary words/style rewrite</td>
<td>49</td>
</tr>
<tr>
<td>Wrong word</td>
<td>102</td>
<td>Wrong tense</td>
<td>46</td>
</tr>
<tr>
<td>Lack of possessive apostrophe</td>
<td>99</td>
<td>Dangling or misplaced modifier</td>
<td>42</td>
</tr>
<tr>
<td>Vague pronoun reference</td>
<td>90</td>
<td>Run-on sentence</td>
<td>39</td>
</tr>
<tr>
<td>No comma in compound sentence</td>
<td>87</td>
<td>Wrong or missing preposition</td>
<td>38</td>
</tr>
<tr>
<td>Pronoun agreement</td>
<td>83</td>
<td>Lack of comma in series</td>
<td>35</td>
</tr>
<tr>
<td>Sentence fragment</td>
<td>82</td>
<td>Its/it’s error</td>
<td>34</td>
</tr>
<tr>
<td>No comma in non-restrictive phrase</td>
<td>75</td>
<td>Tense shift</td>
<td>31</td>
</tr>
</tbody>
</table>
Some Examples

• Comma splice:
  – It is nearly noon, we must stop for food.

• No comma in non-restrictive phrase:
  – The man who I knew well was unhappy.

• Unnecessary comma with restrictive phrase:
  – The man, who I knew well, was unhappy.

• Dangling or misplaced modifier:
  – Turning the corner, a handsome school building appeared.
# Statistics from 3000 Papers

## Error or Error Pattern

<table>
<thead>
<tr>
<th>Error or Error Pattern</th>
<th># found in 3000 papers</th>
<th>% of total errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No comma after introductory element</td>
<td>3,299</td>
<td>11.5%</td>
</tr>
<tr>
<td>2. Vague pronoun reference</td>
<td>2,809</td>
<td>9.8%</td>
</tr>
<tr>
<td>3. No comma in compound sentence</td>
<td>2,446</td>
<td>8.6%</td>
</tr>
<tr>
<td>4. Wrong word</td>
<td>2,217</td>
<td>7.8%</td>
</tr>
<tr>
<td>5. No comma in non-restrictive element</td>
<td>1,864</td>
<td>6.5%</td>
</tr>
<tr>
<td>6. Wrong/missing inflected endings</td>
<td>1,679</td>
<td>5.9%</td>
</tr>
<tr>
<td>7. Wrong or missing preposition</td>
<td>1,580</td>
<td>5.5%</td>
</tr>
<tr>
<td>8. Comma splice</td>
<td>1,565</td>
<td>5.5%</td>
</tr>
<tr>
<td>9. Possessive apostrophe error</td>
<td>1,458</td>
<td>5.1%</td>
</tr>
<tr>
<td>10. Tense shift</td>
<td>1,453</td>
<td>5.1%</td>
</tr>
<tr>
<td>11. Unnecessary shift in person</td>
<td>1,347</td>
<td>4.7%</td>
</tr>
<tr>
<td>12. Sentence fragment</td>
<td>1,217</td>
<td>4.2%</td>
</tr>
<tr>
<td>13. Wrong tense or verb form</td>
<td>952</td>
<td>3.3%</td>
</tr>
<tr>
<td>14. Subject-verb agreement</td>
<td>909</td>
<td>3.2%</td>
</tr>
<tr>
<td>15. Lack of comma in series</td>
<td>781</td>
<td>2.7%</td>
</tr>
<tr>
<td>16. Pronoun agreement error</td>
<td>752</td>
<td>2.6%</td>
</tr>
<tr>
<td>17. Unnecessary comma with restrictive element</td>
<td>693</td>
<td>2.4%</td>
</tr>
<tr>
<td>18. Run-on or fused sentence</td>
<td>681</td>
<td>2.4%</td>
</tr>
<tr>
<td>19. Dangling or misplaced modifier</td>
<td>577</td>
<td>2.0%</td>
</tr>
<tr>
<td>20. Its/it's error</td>
<td>292</td>
<td>1.0%</td>
</tr>
</tbody>
</table>
Key Findings

• The distribution of errors across types changes over time
• The absolute proportion of errors in text appears remarkably consistent over time
• At least in this study, many errors appear to be caused by ‘declining familiarity with the visual look of a written page’ — the impact of an oral culture?
Other Taxonomies of Error #1: Douglas and Dale 1991

- Spelling Errors
- Syntactic Errors
  - Semantic Problems
  - Stylistic Problems
  - Rhetorical Problems
  - Punctuation Problems
- Co-occurrence Errors
- Dependency and Subordination Errors
  - Number Disagreement
  - Bad Subcategorisation
  - Resumptive Pronoun
  - Syntactic Parallelism
  - Bad Clause Conjunction
  - Misleading PP Attachment
  - Misleading Adverbial Attachment
  - Missing Subordination Indicator
  - Redundancy
Other Taxonomies of Error #1: Douglas and Dale 1991

- Spelling Errors
- Syntactic Errors
- Semantic Problems
- Stylistic Problems
- Rhetorical Problems
- Punctuation Problems
- Wrong Lexical Choice
- Wrong Lexical-Semantic Combination
- Semantic Redundancy
- Obscure Pronominal Antecedents
- Intended Sense Failure
- Semantic Parallelism Failure
Other Taxonomies of Error #1: Douglas and Dale 1991

- Spelling Errors
- Syntactic Errors
- Semantic Problems
- Stylistic Problems
- Rhetorical Problems
- Punctuation Problems

- Non-preferred Words
- Vague and Redundant Phrases
- Too Many Embedded PPs
- Too Many Subordinate Clauses
- Excessive Nominalisation
- Unnecessary Subjunctive
- Unnecessary Clumsy Passive
- Clumsy Constructions
- House Style Errors
Other Taxonomies of Error #1:
Douglas and Dale 1991

- Spelling Errors
- Syntactic Errors
- Semantic Problems
- Stylistic Problems
- Rhetorical Problems
- Punctuation Problems
- Delimiter Problems
- Comma as List Separator
- Punctuation Hierarchy Problems
- Unnecessary Commas
- Punctuation Placement Errors
- Apostrophe Misuse
- Capitalisation Errors
- Spacing Errors
Other Taxonomies of Error #2: Becker et al 2003

- Morphological errors
- Syntax errors: word order, categorial, case, and agreement errors
- Syntactic–semantic selection (such as for fixed verbal structures)
- Orthographic errors
- Syntax errors which cannot be easily classified further
- Competence errors which can’t easily be reconstructed (corrected) to form grammatical sentences
Other Taxonomies of Error #3: Bušta 2009

- Spelling
- Morphology
- Syntax
- Punctuation
- Lexical and semantic choice
- Style
- Typography
**Other Taxonomies of Error #2: Bušta 2009**

<table>
<thead>
<tr>
<th>Error Group</th>
<th>count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling (simple)</td>
<td>2,347</td>
<td>13.04</td>
</tr>
<tr>
<td>Morpho-syntactic</td>
<td>1,689</td>
<td>9.39</td>
</tr>
<tr>
<td>Spelling (other)</td>
<td>867</td>
<td>4.82</td>
</tr>
<tr>
<td>Lexico-semantics</td>
<td>2,536</td>
<td>14.09</td>
</tr>
<tr>
<td>Punctuation</td>
<td>3,837</td>
<td>21.32</td>
</tr>
<tr>
<td>Stylistic</td>
<td>4,184</td>
<td>23.25</td>
</tr>
<tr>
<td>Typography</td>
<td>2,165</td>
<td>12.03</td>
</tr>
<tr>
<td>unsorted</td>
<td>371</td>
<td>2.06</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>17,996</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*Fig. 5. Error classification group statistics in the Chyby corpus.*
Prevailing Findings

• A large proportion of errors are very simple
• The nature of the errors to be dealt with depend on the context of writing production:
  – Non-native speakers
  – Authored text being copyedited
  – Technical manuals
  – Translations
• But: maybe you can’t see the wood for the trees—complex errors may be ignored or considered out of scope
Complex Errors

• The living area is something you would expect to find in a house, let alone an apartment.

• If there are mistakes to be acknowledged, we will not shy away from doing so.

• How can one write a minimal manual, not as a cut-down version of a conventional manual, but derived from first principles of what users need successfully to start up their use of a system, and to provide the basis of their subsequent learning of it?
Conclusions

• Many problems in writing are what we might think of as ‘low level’ errors: spelling, punctuation, typographic mistakes . . .

• . . . but many of the problems in real texts are at a higher level than straightforward textbook grammar errors

• Many problems of both types can be characterised as cases where the essence is correct but the rendering is incorrect