CENTRE FOR LANGUAGE TECHNOLOGY
MACQUARIE UNIVERSITY ~ SYDNEY

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

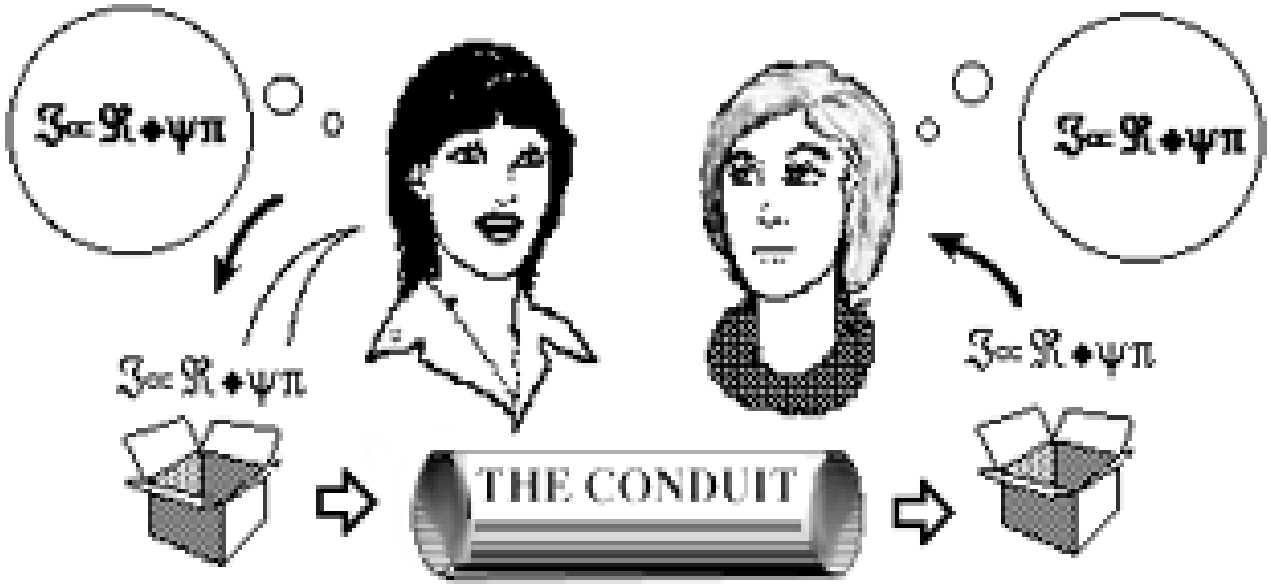
What This Course is Not About

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

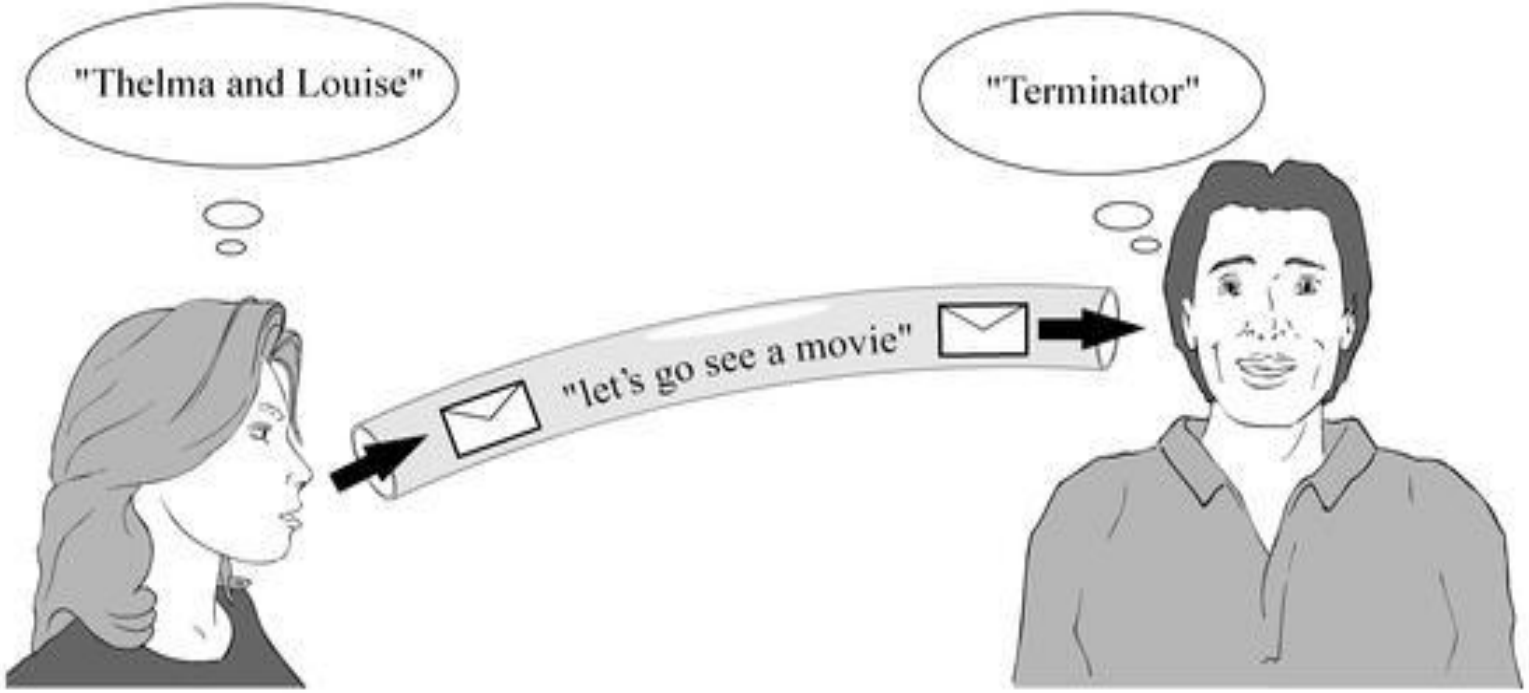
Overview

- The Nature of the Writing Process
- Categorising Errors

The Conduit Metaphor #1



The Conduit Metaphor #2



Kinds of Writing

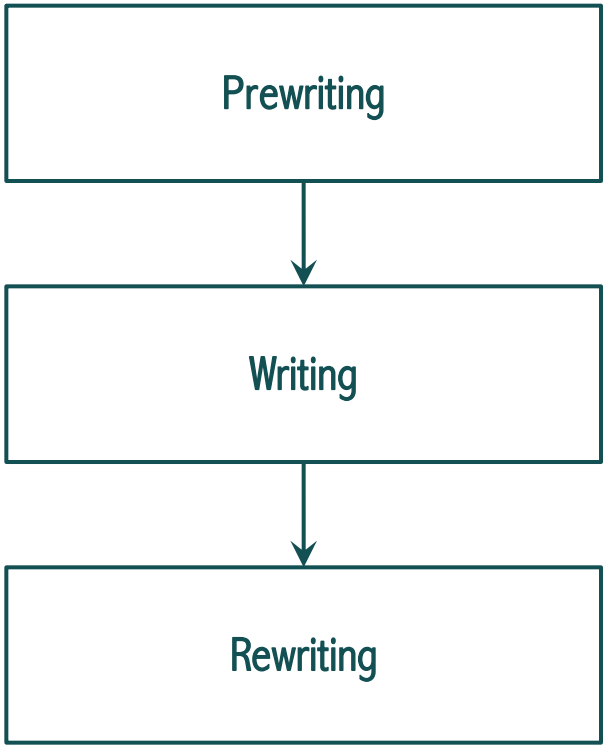
- Fiction
- Poetry
- Personal journals or narratives
- News reporting
- Technical writing
- …

Technical Writing

… 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.

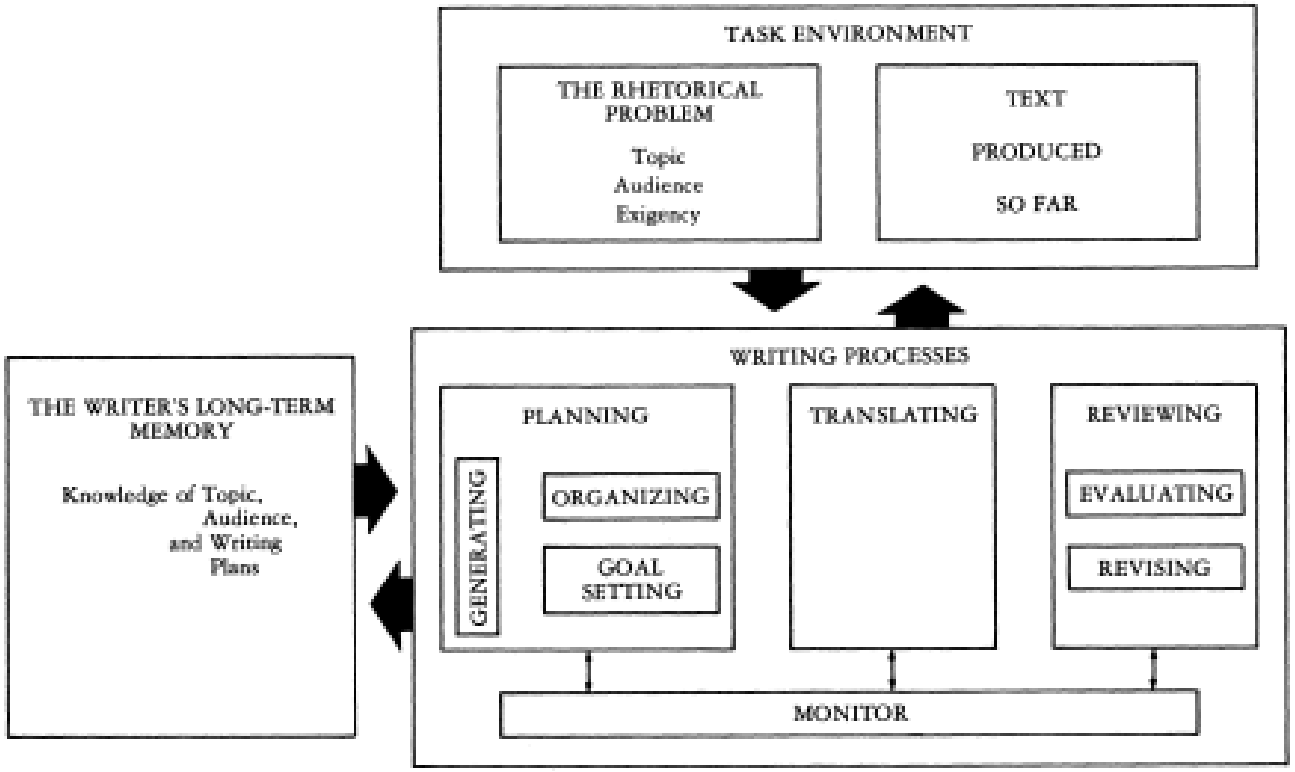
Britton 1965:114

A Stage Model of the Writing Process



Rohman 1965

A Cognitive Process Model



Flower and Hayes 1981

The Nature of Revision

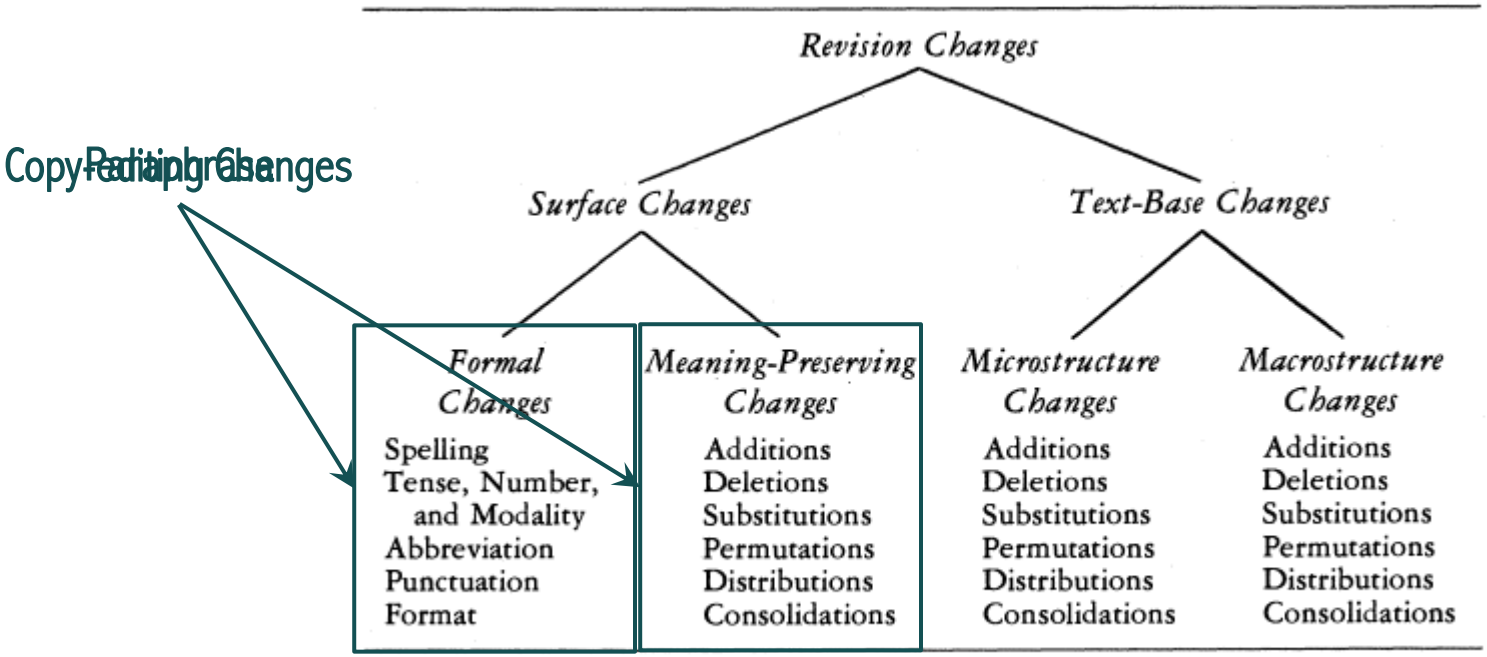


Figure 1. A Taxonomy of Revision Changes

Faigley and Witte 1981

Meaning-Preserving Changes: Additions

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

Meaning-Preserving Changes: Substitutions

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

Meaning-Preserving Changes: Distributions

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.

Meaning-Preserving Changes: Consolidations

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

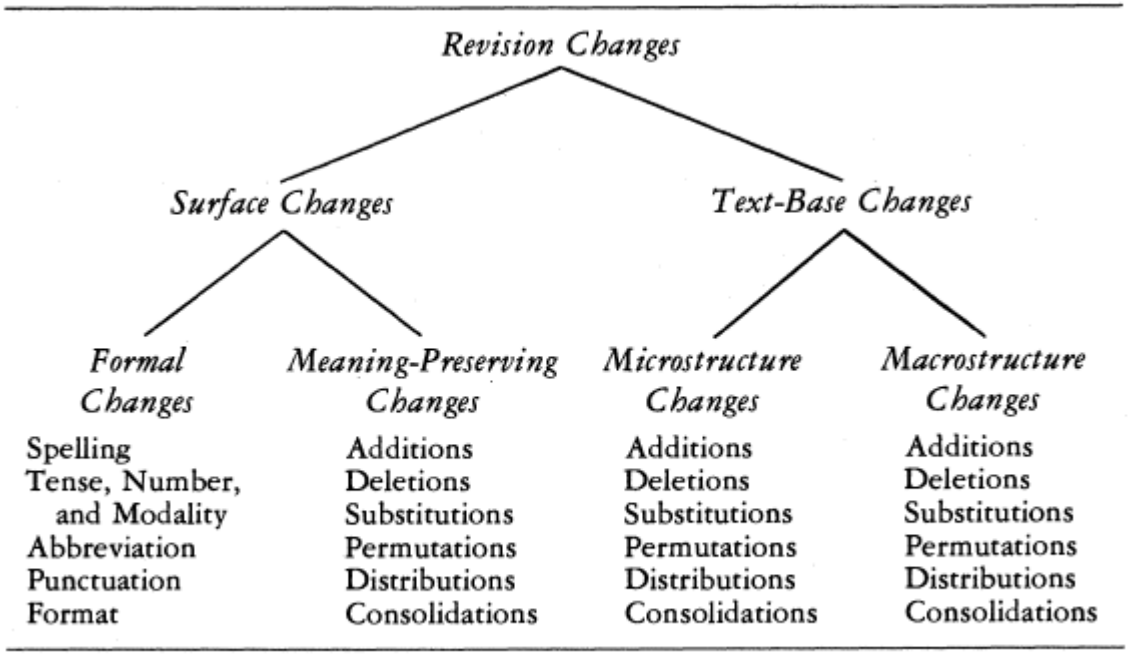


Figure 1. A Taxonomy of Revision Changes

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

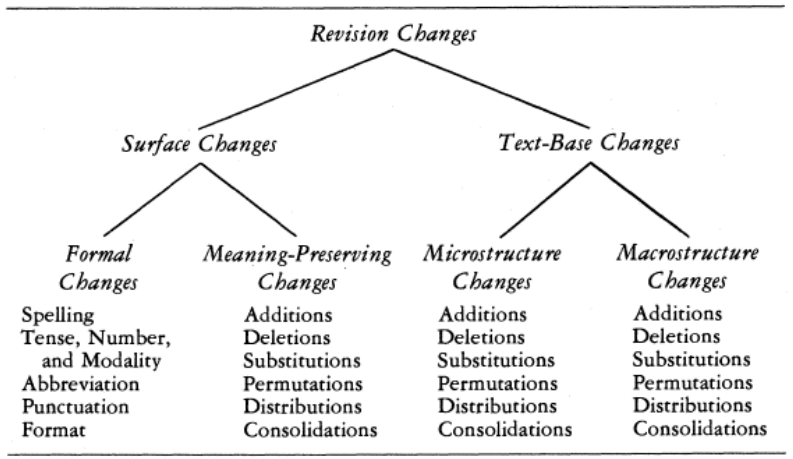
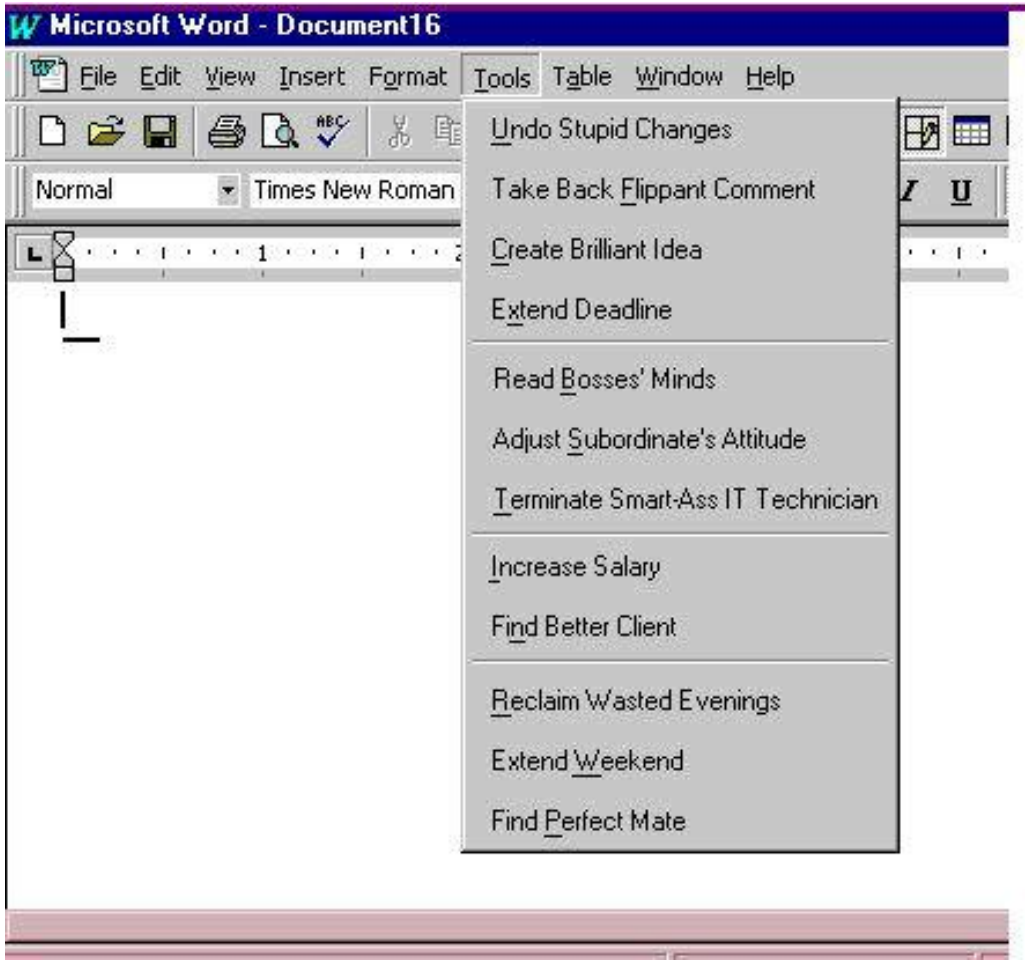


Figure 1. A Taxonomy of Revision Changes

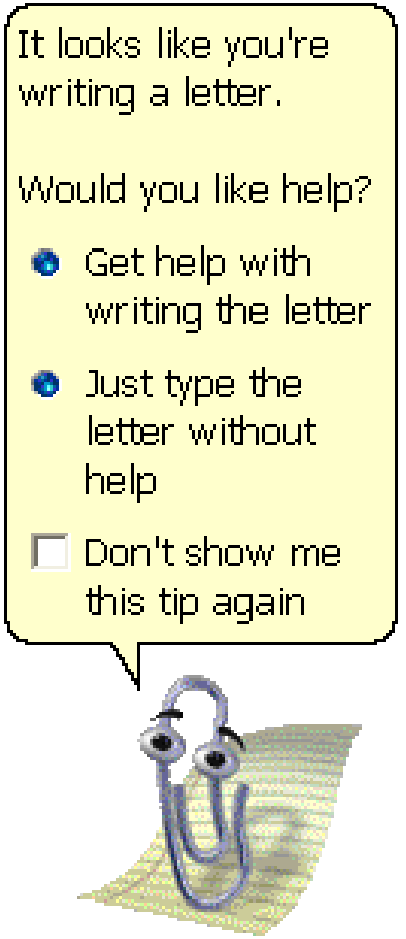
- We can also conceive of machine assistance being provided for the prewriting and writing stages

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**

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

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.

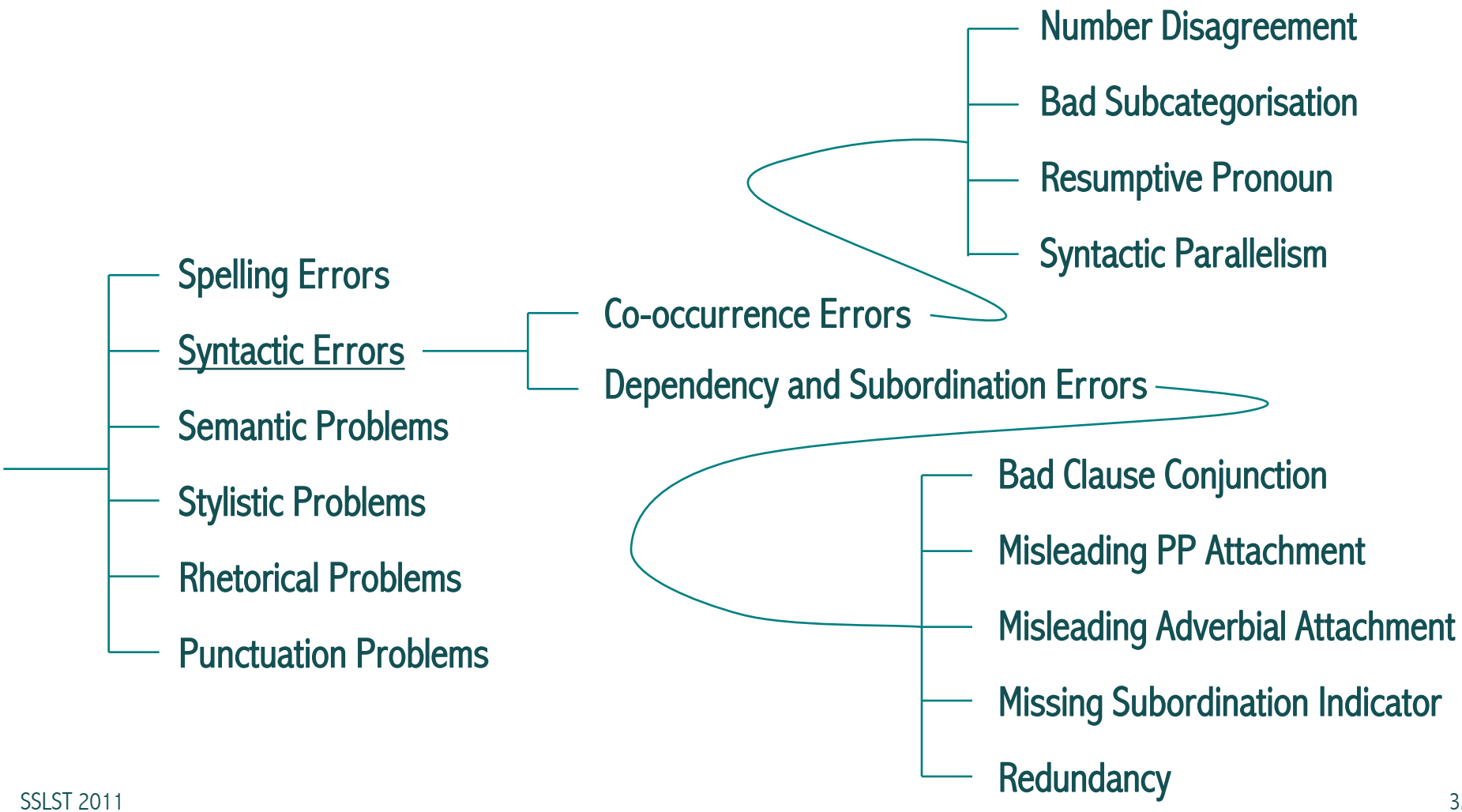
Statistics from 3000 Papers

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

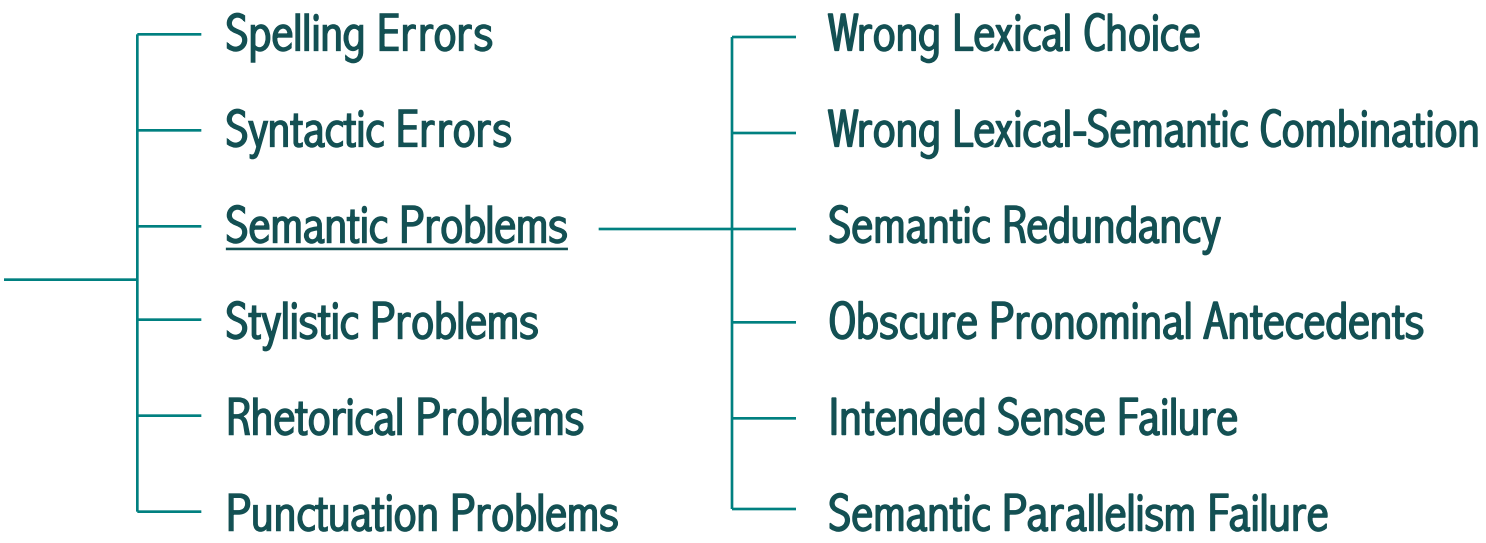
Key Findings

- 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?

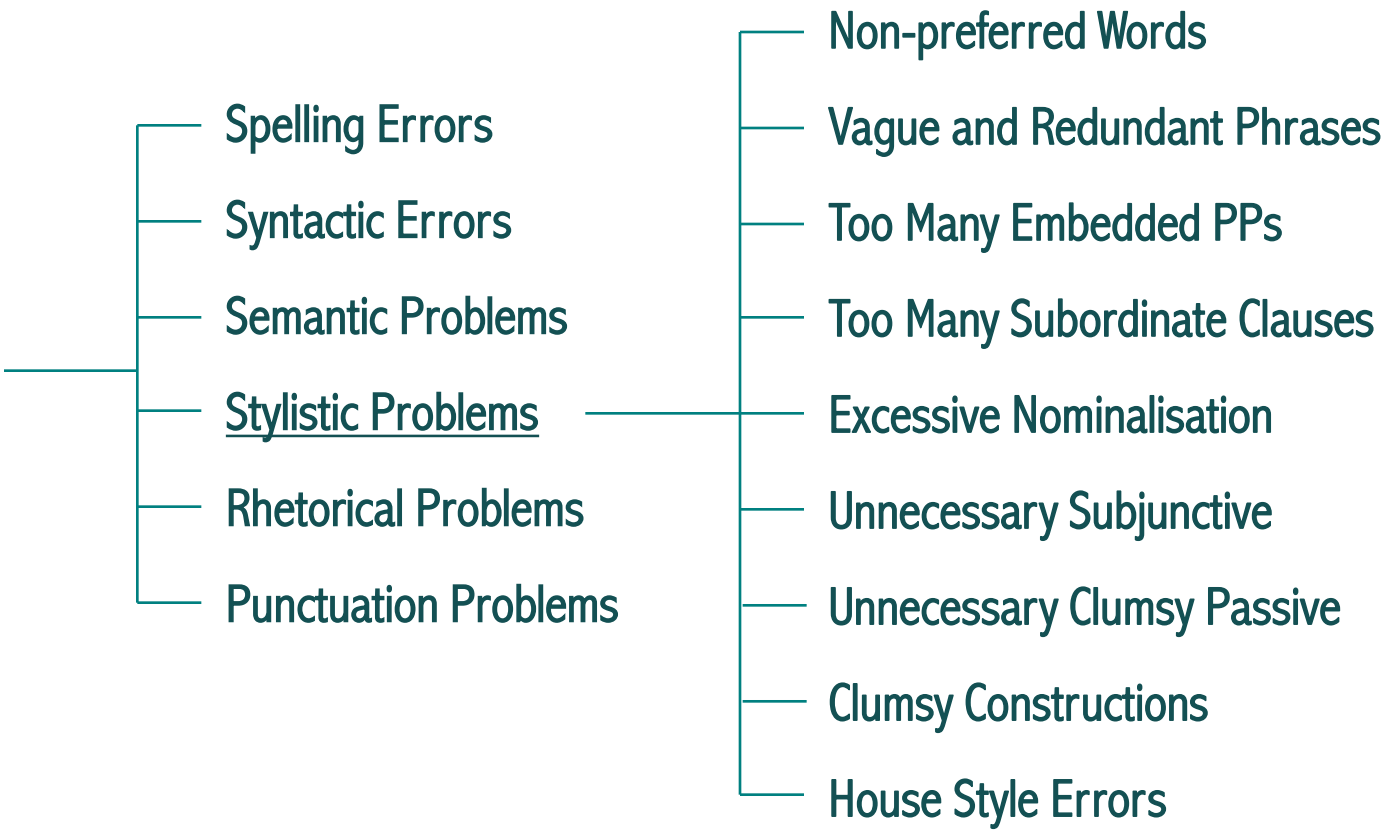
Other Taxonomies of Error #1: Douglas and Dale 1991



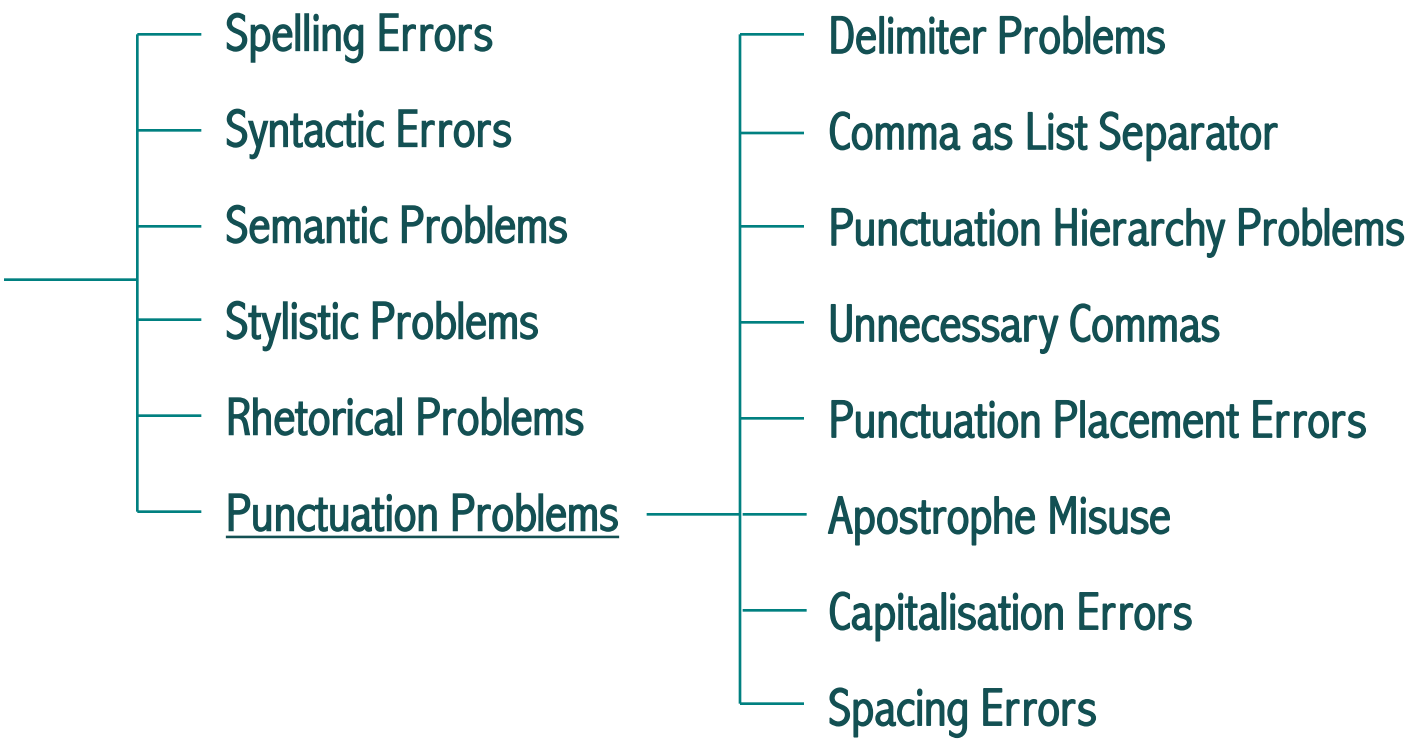
Other Taxonomies of Error #1: Douglas and Dale 1991



Other Taxonomies of Error #1: Douglas and Dale 1991



Other Taxonomies of Error #1: Douglas and Dale 1991



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

Error Group	count	%
Spelling (simple)	2,347	13.04
Morpho-syntactic	1,689	9.39
Spelling (other)	867	4.82
Lexico-semantics	2,536	14.09
Punctuation	3,837	21.32
Stylistic	4,184	23.25
Typography	2,165	12.03
unsorted	371	2.06
Total	**17,996**	100.00

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