Helping People Write: Grammar Checking and Beyond

Robert Dale
Centre for Language Technology
Macquarie University

What This Course is About

- How we can use NLP tools and techniques to help people write:
  - Spell checking
  - Grammar checking
  - Style checking and discourse-level assistance
- These materials, along with a bibliography, are available at:
  - http://web.science.mq.edu.au/~rdale/teaching/icon2011

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

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

Common Errors in English

Ways of Categorising Errors

- By symptom:
  - Misspelled words
  - Ungrammatical sentences
  - Stylistic disfluencies and inconsistencies
- By cause:
  - Mechanical errors [also known as errors of execution or performance errors]
  - Cognitive errors [also known as errors of intention or competence errors]

An Analysis of Student Writing Errors

- Connors and Lunsford 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%

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: 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 problems in real texts are at a higher level than straightforward textbook grammar errors

Overview

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

What is a Spelling Error?

- How many spelling errors are there here?
  - Wot color is the dawg?
  - C u l8er
- A definition:
  - A spelling error is a word which is not spelled as it should be

Use Cases for Spell Checking

- Correcting spelling errors in text
- Fixing OCR output
- Correcting spelling errors in search queries
- Some solutions allow interaction, others require machine autonomy

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

Non-Word Errors vs Real-Word Errors

- The boys ate thier toast.
- The boys ate there toast.

Unix Spell

```
$ spell
This is the storry abuot an event that went from
baad to wurse
abuot
baad
storry
wurse
$
```

Storage Issues

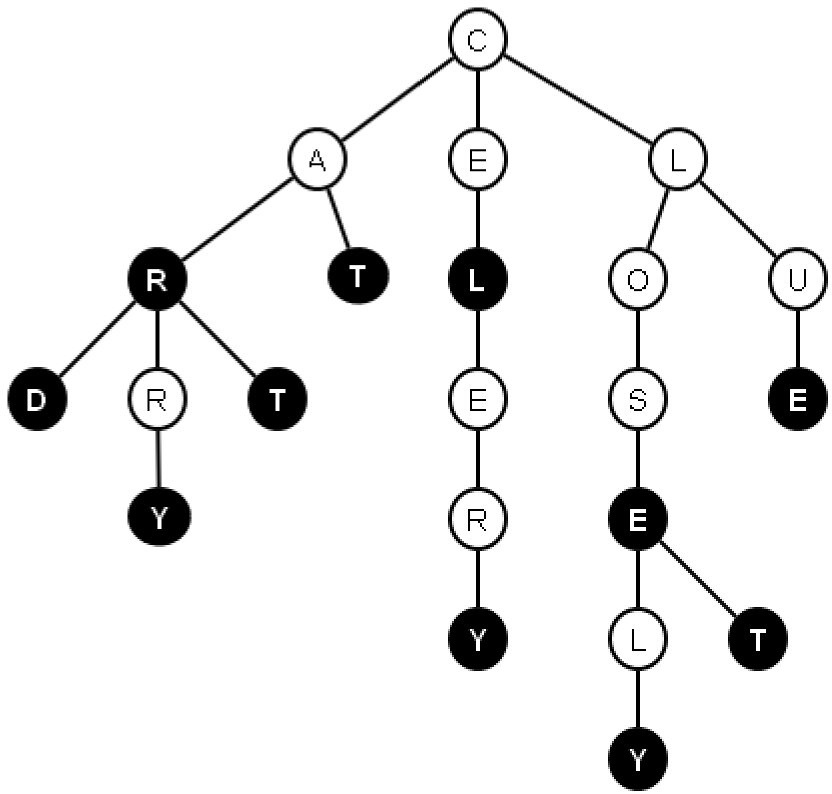
1981: The original PC's maximum memory using IBM parts was 256 KB: 64 KB on the motherboard and three 64 KB expansion cards.

A word list of100k words occupies around 500KB.

Peterson's Three Levels of Storage

- Small dictionary of frequently used words [100–200 words]
- Document-specific words [1000–2000 words]
- Larger secondary storage [10k–100k words]

Dictionary Storage via Tries



Problems with Word Lists

- False Negatives
  - A misspelled word may not be flagged as a spelling error because it is orthographically identical to some other valid word
- False Positives
  - A valid word may be flagged as a spelling error because it is not in the list

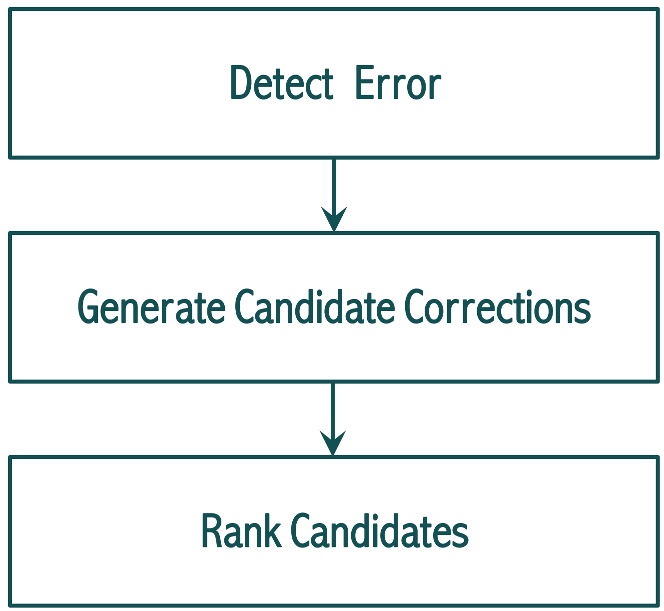
Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

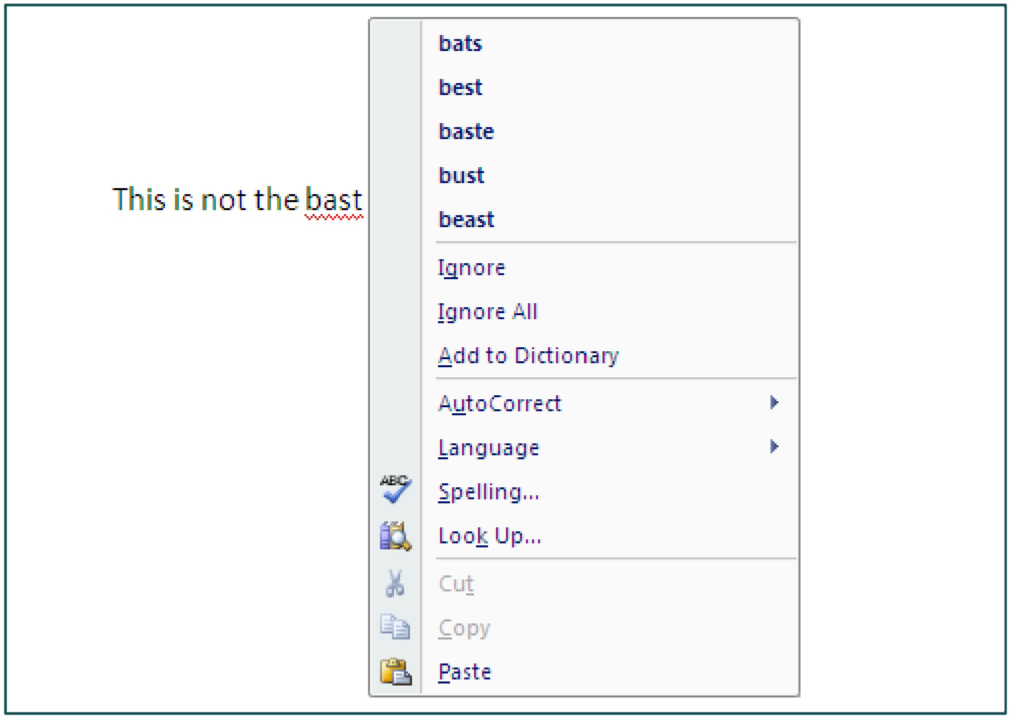
The Task

- Given a word which is assumed to be misspelled, find the word that the author intended to type

Spell Checking



Interactive Correction with Candidates



Finding Candidate Corrections

- Look for ‘nearby’ real words
- Edit distance:
  - An edit = a deletion, an insertion, a transposition or a substitution
  - Each edit adds 1 to the edit distance between strings
- Damerau 1980: 80% of spelling errors are 1 edit from the correct string

Edit Distance

- Deletion:
  - continous → continuous
- Insertion:
  - explaination → explanation
- Substitution
  - anyboby → anybody
- Transposition:
  - autoamtically → automatically

Using Edit Distance

- For a hypothesized misspelled word:
  - Generate all strings within an edit distance of 1
  - Filter non-words out of the list

teh	$\rightarrow$	tea	$\rightarrow$	tea
		teb		~~teb~~
		...		...
		the		the

Potential Problems with Edit Distance

- For a string of n characters from an alphabet of size k, number of strings within edit distance 1:

$$k(2n+1)+n-1$$

- Peterson [1980]: an average of 200 dictionary accesses for each misspelling
- Also: words > edit distance 1 are ignored

Approaches to Spelling Correction

- Yannakoudakis and Fawthrop [1983]: Error Patterns
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

What Causes Spelling Errors?

- Typing errors (typographic errors, errors of execution)

  the $\rightarrow$ teh

  spell $\rightarrow$ speel

- Cognitive errors (orthographic errors, errors of intention)

  receive $\rightarrow$ recieve

  conspiracy $\rightarrow$ conspiricy

  abyss $\rightarrow$ abiss

  naturally $\rightarrow$ nacherly

Yannakoudakis and Fawthrop [1983]: Error Patterns

- Problem Statement:
  - Given a non-word error, generate a ranked list of candidate replacements based on common error patterns
- Background assumption:
  - Many errors are due to phonetic confusion
  - But conversion into a phonetic coding assumes a dialect

Yannakoudakis and Fawthrop [1983]: The Approach

- Analysed a corpus of 1377 spelling errors
- Divide each word into spelling elements – a bit like vowel and consonant clusters, but oriented towards typical confusions in spelling:
  - F-OR-EI-GN
  - D-I-PH-TH-ER-IA
  - F-A-V-OUR-A-B-L-E

Yannakoudakis and Fawthrop [1983]: Error Rules

- A ‘vocabulary’ of 299 spelling elements
- Very large space of possible element-to-element replacements
- Constrained by observed patterns:
  - Doubling or singling of characters
  - Errors involving specific characters
  - Errors involving related phonemes
  - …

→3079 error rules

Yannakoudakis and Fawthrop [1983]: Other Heuristics

- The most frequent length of an error form is one character less than the dictionary form
- Typing errors are caused by hitting an adjacent key to the one intended or by hitting the correct key and its neighbour
- Short error forms do not contain more than one error
- If the error form is short, only dictionary words differing in length by one character from the error form are examined

Yannakoudakis and Fawthrop [1983]: Examples

- F-ILIPIN-OE-S → PH-ILIPIN-O-S
- CA-PH-EINE → CA-FF-EINE
- When there's more than one possible correction, choice is made via 'subjective Bayesian probabilities' on the dictionary words and the error rules

Yannakoudakis and Fawthrop [1983]: Performance

- Corrected 90% of 1153 error forms
  - In 95% of these corrections one word was identified
  - In 5% a choice of between 2 and 4 words was offered
- Mean time to correct an error was 22 seconds, with a minimum of five seconds and a maximum of 50 seconds

Approaches to Spelling Correction

- Yannakoudakis and Fawthrop [1983]: Error Patterns
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Kernighan, Church and Gale [1990]: Using the Noisy Channel Model

- The problem:
  - Given a word in error, find the most likely word intended by the author
- Approach:
  - Find all words within edit distance of 1
  - Determine the probability of each possible <u>edit</u> from a corpus
  - Use these probabilities to order the list of candidates

Kernighan, Church and Gale [1990]: An Example: Candidate Corrections

Typo	Correction	Transformation	
acress	actress	@ t 2	deletion
acress	cress	a # 0	insertion
acress	caress	ac ca 0	reversal
acress	access	r c 2	substitution
acress	across	e o 3	substitution
acress	acres	s # 4	insertion
acress	acres	s # 5	insertion

Kernighan, Church and Gale [1990]: Using the Noisy Channel Model

- We want to find the most likely correction c given a misspelling t
- By Bayes Rule, this means finding the c that maximizes

$$Pr(c).Pr(t|c)$$

Prior model of word probabilities

The channel (or error) model

Kernighan, Church and Gale [1990]: Prior Probabilities

- Pr(c) is estimated by:

$$\frac{freq(c) + 0.5}{N}$$

- where freq(c) is the number of times that the word c appears in the 1988 AP corpus (N = 44 million words)

Kernighan, Church and Gale [1990]: Conditional Probabilities

$$Pr(t|c) \approx \begin{cases} \dfrac{del[c_{p-1}, c_p]}{chars[c_{p-1}, c_p]}, & \text{if deletion} \\ \dfrac{add[c_{p-1}, t_p]}{chars[c_{p-1}]}, & \text{if insertion} \\ \dfrac{sub[t_p, c_p]}{chars[c_p]}, & \text{if substitution} \\ \dfrac{rev[c_p, c_{p+1}]}{chars[c_p, c_{p+1}]}, & \text{if reversal} \end{cases}$$

- *del*, *add*, *sub* and *rev* are derived from confusion matrices
- *chars* are occurrence counts derived from the corpus

Kernighan, Church and Gale [1990]: Confusion Matrices

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Kernighan, Church and Gale [1990]: The Example: Scoring the Candidates

Correction	Score	Raw	freq(c)	Pr(t|c)
actress	37%	.157	1343	55/470,000
cress	0%	.000	0	46/32,000,000
caress	0%	.000	4	0.95/580,000
access	0%	.000	2280	0.98/4,700,000
across	18%	.077	8436	93/10,000,000
acres	21%	.092	2879	417/13,000,000
acres	23%	.098	2879	205/6,000,000

Kernighan, Church and Gale [1990]: The Example in Context

… was called a "stellar and versatile <u>acress</u> whose combination of sass and glamour has defined her ….

Kernighan, Church and Gale [1990]: Performance

- Test sample of 329 misspelled words with two candidate corrections
- Program agrees with majority of judges in 87% of cases

Approaches to Spelling Correction

- Yannakoudakis and Fawthrop [1983]: Error Patterns
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Agirre et al [1998]: Using Context

- The Goal:
  - Given a non-word error, use the <u>context</u> to determine the most likely correction (the ‘single proposal’)

Agirre et al [1998]: Sources of Knowledge

- Statistical:
  - Brown Corpus (BF) and document (DF) word frequencies
- Syntactic:
  - Constraint Grammar (CG) used to rule out candidate corrections that are grammatically unacceptable
- Semantic:
  - Use distance in WordNet (CD) to choose the candidate noun correction that is closest to the words in the context

Agirre et al [1998]: Performance

- A large number of combinations tried on artificially generated error data
- Best performing combinations tested on real error data
- Main findings:
  – Combination of syntax and document frequencies works best
    – But effect of document frequency impacted by small documents
  – Brown Corpus frequencies and conceptual density not useful

Approaches to Spelling Correction

- Yannakoudakis and Fawthrop [1983]: Error Patterns
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Brill and Moore [2000]: Improving the Noisy Channel Model

- The Approach:
  - Given a word assumed to be in error, use a noisy channel model based on <u>string to string edits</u> to determine candidate corrections

Brill and Moore [2000]: Approach

- Generalise the error model to permit generic string to string edits

  $\Pr(\alpha \rightarrow \beta)$ is the probability that the user types β when they meant α

- Edits are conditioned on position in the string:
  - $\Pr(\alpha \rightarrow \beta \mid PSN)$ where PSN = start, middle, or end of word
- Observation:
  - $P(e \mid a)$ does not vary by location
  - $P(ent \mid ant)$ does

Brill and Moore [2000]: Example

- Spelling error:
  - physical $\rightarrow$ fisikle
- Conceptually, the user picks a word; partitions it into substrings; generates each partition, perhaps erroneously
  - ph+y+s+i+c+al $\rightarrow$ f+i+s+i+k+le
- Probability of generating the error is then:
  - $P(f \mid ph) \,.\, P(i \mid y) \,.\, P(s \mid s) \,.\, P(i \mid i) \,.\, P(k \mid c) \,.\, P(le \mid al)$

Brill and Moore [2000]: Learning the Model

- String to string edits are derived from mismatches in aligned ⟨spelling error, correction⟩ pairs:

```
  A C T U A L
A K G S U A L
```

- Edits derived:

  c → k, ac → ak, c → kg, ac → akg, ct → kgs

Brill and Moore [2000]: Testing

- 10000 word corpus of spelling errors + corrections
- 200k word dictionary
- Language model assigns uniform probabilities to all words

Brill and Moore [2000]: Performance

Without positional information:

Max Window	1-Best	2-Best	3-Best
0	87.0	93.9	95.9
Church and Gale	89.5	94.9	96.5
1	90.9	95.6	96.8
2	92.9	97.1	98.1
3	93.6	97.4	98.5
4	93.6	97.4	98.5

Brill and Moore [2000]: Performance

With positional information:

Max Window	1-Best	2-Best	3-Best
0	88.7	95.1	96.6
1	92.8	96.5	97.4
2	94.6	98.0	98.7
3	95.0	98.0	98.8
4	95.0	98.0	98.8
5	95.1	98.0	98.8

Approaches to Spelling Correction

- Yannakoudakis and Fawthrop [1983]: Error Patterns
- Kernighan, Church and Gale [1990]: The Noisy Channel Model
- Agirre et al [1998]: Using Context
- Brill and Moore [2000]: String-to-String Edits
- Toutanova and Moore [2002]: Pronunciation Modeling

Toutanova and Moore [2002]: Pronunciation Modeling

- Observation:
  - Many errors in Brill and Moore [2000] are due to word pronunciation

Misspelling	Correct Word	B+M Proposal
edelvise	edelweiss	advice
bouncie	bouncy	bounce
latecks	latex	lacks

Toutanova and Moore [2002]: Approach

- Build <u>two</u> error models:
  - The Brill and Moore [2000] model
  - A phone-sequence to phone-sequence error model
- Uses machine-learned letter-to-phone conversion
- At classification time, the two models are combined using a log linear model

Toutanova and Moore [2002]: Performance

Model	1-Best	2-Best	3-Best	4-Best
Brill and Moore	94.21	98.18	98.90	99.06
Phoneme	86.36	93.65	95.69	96.63
Combined	95.58	98.90	99.34	99.5
Error Reduction	23.8	39.6	40	46.8

Toutanova and Moore [2002]: Examples

Misspelling	Correct	LTR Guess
bouncie	bouncy	bounce
edelvise	edelweiss	advise
grissel	gristle	grizzle
latecks	latex	lacks
neut	newt	nut
rench	wrench	ranch
saing	saying	sang
stail	stale	stall

Spell Checking

- What's a Spelling Error?
- Non-Word Error Detection
- Error Correction
- Real-Word Error Detection

Real Word Errors are a Real World Problem

- Peterson:
  - 10% of typing errors are undetected when using a 50k word dictionary
  - 15% are undetected when using a 350k word dictionary
- Two Main Approaches in the Literature:
  1. Try to determine from contextual evidence whether a word is a real-word error
  2. Given a potential real-word error, determine the most likely correction

Mays, Damerau and Mercer [1991]: Using Trigrams to Detect Real-Word Errors

- The Goal:
  - Given a text, determine presence of real-word errors and propose candidate corrections
- Basic Idea:
  - If the trigram-derived probability of an observed sentence is lower than that of any sentence obtained by replacing one of the words with a spelling variation, then hypothesize that the original is an error and the variation is what the user intended.

Mays, Damerau and Mercer [1991]: The Idea

- Example:
  - I saw the man it the park
- Syntax can be used:
  - to determine that an error is present
  - to determine whether candidate corrections result in grammatical strings
- But we don't have 100% reliable parsers, so try something else: a trigram language model

Mays, Damerau and Mercer [1991]: The Key Insights

- A low-probability word sequence can be considered evidence for a real-word error
- High-probability sequences can be used to rank correction candidates

Mays, Damerau and Mercer [1991]: The Data

- Restricted to edit distance 1 errors, and one misspelled word per sentence
- Given a set of 100 randomly selected sentences:
  - For each sentence, generate all possible sentences where each word is subjected to edit distance 1 transformations

→ 8628 misspelled sentences

Mays, Damerau and Mercer [1991]: The Noisy Channel Model

- We want to find the most likely correction w given a misspelling x
- By Bayes Rule, this means finding the w that maximizes

$$Pr(w).Pr(x|w)$$

The channel model

Prior model of word probabilities, approximated using the trigram model

Mays, Damerau and Mercer [1991]: The Noisy Channel Model

- The channel model:

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ (1-\alpha)/|SV(w)| & \text{if } x \in SV(w) \\ 0 & \text{otherwise} \end{cases}$$

- SV(w) is the set of spelling variations of w; all are considered equally likely
- The challenge: find the optimal value for α, the a priori belief that the observed input word is correct

Mays, Damerau and Mercer [1991]: Performance

α	Original	Changed	Correct	Composite
0.9000	15.0	94.4	78.7	74.4
0.9900	3.0	86.9	90.9	79.0
0.9990	1.0	76.7	95.4	73.2
0.9999	0.0	63.7	97.0	61.8

- original = %age of original input sentences changed to some other sentence
- changed = %age of misspelled sentences changed to some other sentence
- correct = %age of changed misspelled sentences that were changed correctly
- composite = %age of misspelled sentences that were changed correctly

Mays, Damerau and Mercer [1991]: Observations

- As α increases the correctness of the changes increases
- As α increases the percentage of misspelled sentences changed to some other sentence decreases
- A reasonable value for α lies in the range 0.99–0.999

See Wilcox-O'Hearn, Hirst and Budanitsky [2008] for a rational reconstruction and proposals for improvements

Hirst and Budanitsky [2005]: Lexical Cohesion for Real-Word Error Correction

- The Goal:
  - Determine real-word errors on the basis of their semantic incompatibility with the rest of the text
- Basic idea:
  - Words which are semantically unrelated to the context, but whose spelling variations <u>are</u> related to the context, are possible real-world spelling errors

Hirst and Budanitsky [2005]: Syntax Doesn’t Always Help

- It is my sincere hole [hope] that you will recover swiftly.
- The committee is now [not] prepared to grant your request.

Hirst and Budanitsky [2005]: The Underlying Observation

- Linguistic cohesion is maintained by lexical chains: words linked by lexical and semantic relationships
  - literal repetition
  - coreference
  - synonymy
  - hyponymy

Hirst and Budanitsky [2005]: Key Assumptions

- A real-word spelling error is unlikely to be semantically related to the text.
- Usually, the writer's intended word will be semantically related to nearby words.
- It is unlikely that an intended word that is semantically unrelated to all those nearby will have a spelling variation that is related.
- So: detect tokens that fit into no lexical chain in the text and replace them with words for which they are plausible mistypings that <u>do</u> fit into a lexical chain.

Hirst and Budanitsky [2005]: Requirements

- A mechanism for generating candidate spelling variations
  - For example, all real words in edit distance 1
- A mechanism for determining whether two words are semantically related
  - For example, distance measures in WordNet

Hirst and Budanitsky [2005]: The Approach

- Ignore words not in the lexicon, closed class words, and elements of a list of non-topical words (eg know, find, world)
- For any remaining suspect:
  - Determine if it is semantically related to another word in the text
  - If not, then look for positive evidence: is any spelling variation a better fit?

Hirst and Budanitsky [2005]: Performance

Scope	Detection		
	P_D	R_D	F_D
1	0.184	0.498	0.254
3	0.205	0.372	0.245
5	0.219	0.322	0.243
MAX	0.247	0.231	0.211
Chance	0.0129	0.0129	0.0129

But: Wilcox-O'Hearn et al [2008] show that the Mays, Damerau, and Mercer model performs better.

Whitelaw et al [2009]:
The Web as a Corpus for Spelling Correction

- Basic idea:
  - Use the web as a large noisy corpus to infer knowledge about misspellings and word usage
  - Avoid using any manually-annotated resources or explicit dictionaries
- Important feature: easily ported to other languages

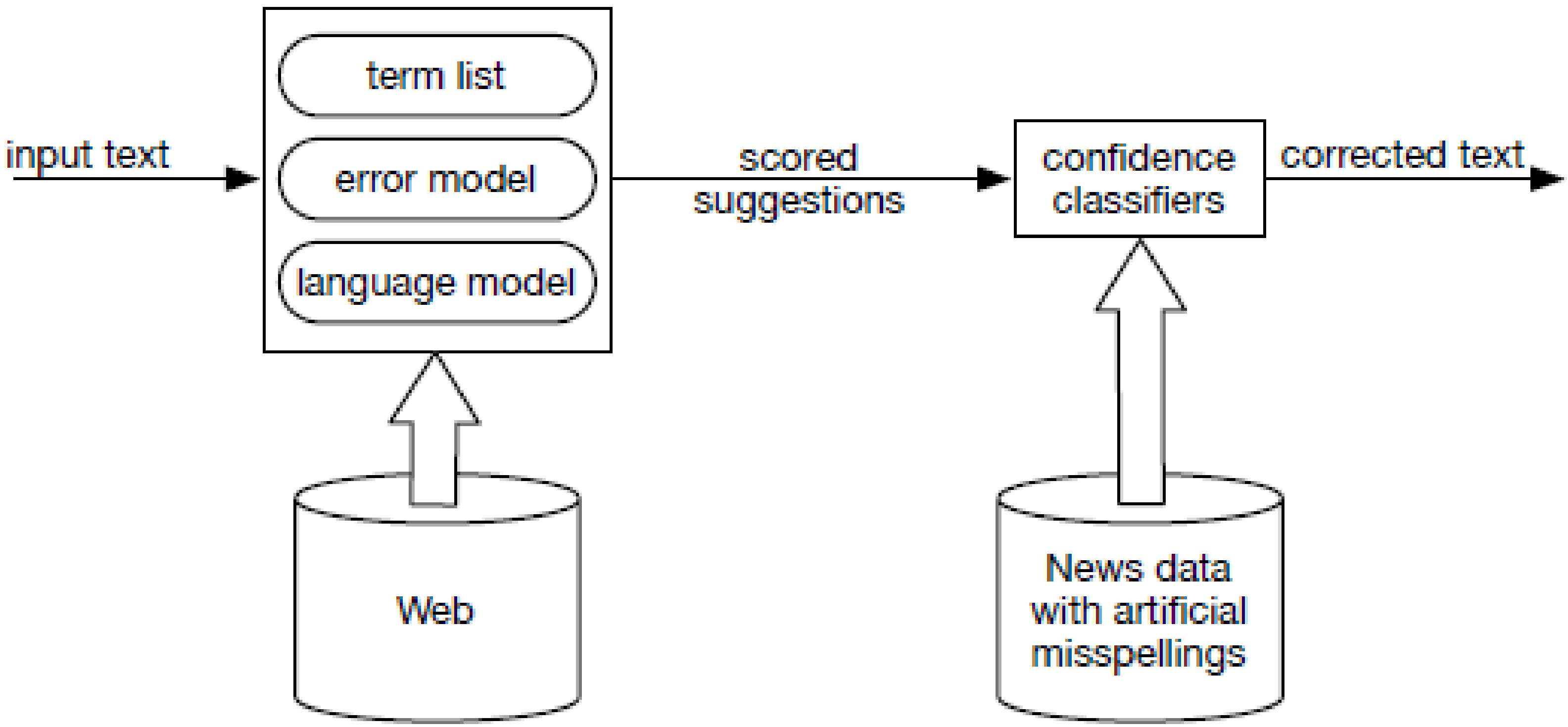
Whitelaw et al [2009]: Approach

- Infer information about misspellings from term usage observed on the Web, and use this to build an error model
- The most frequently observed terms are taken as a noisy list of potential candidate corrections
- Token n-grams are used to build a language model which is used to make context-appropriate corrections

Whitelaw et al [2009]: Key Feature

- Given error and LM scores, confidence classifiers determine the thresholds for spelling error detection and auto-correction
- Classifiers are trained on clean news data injected with artificial misspellings

Whitelaw et al [2009]: System Architecture



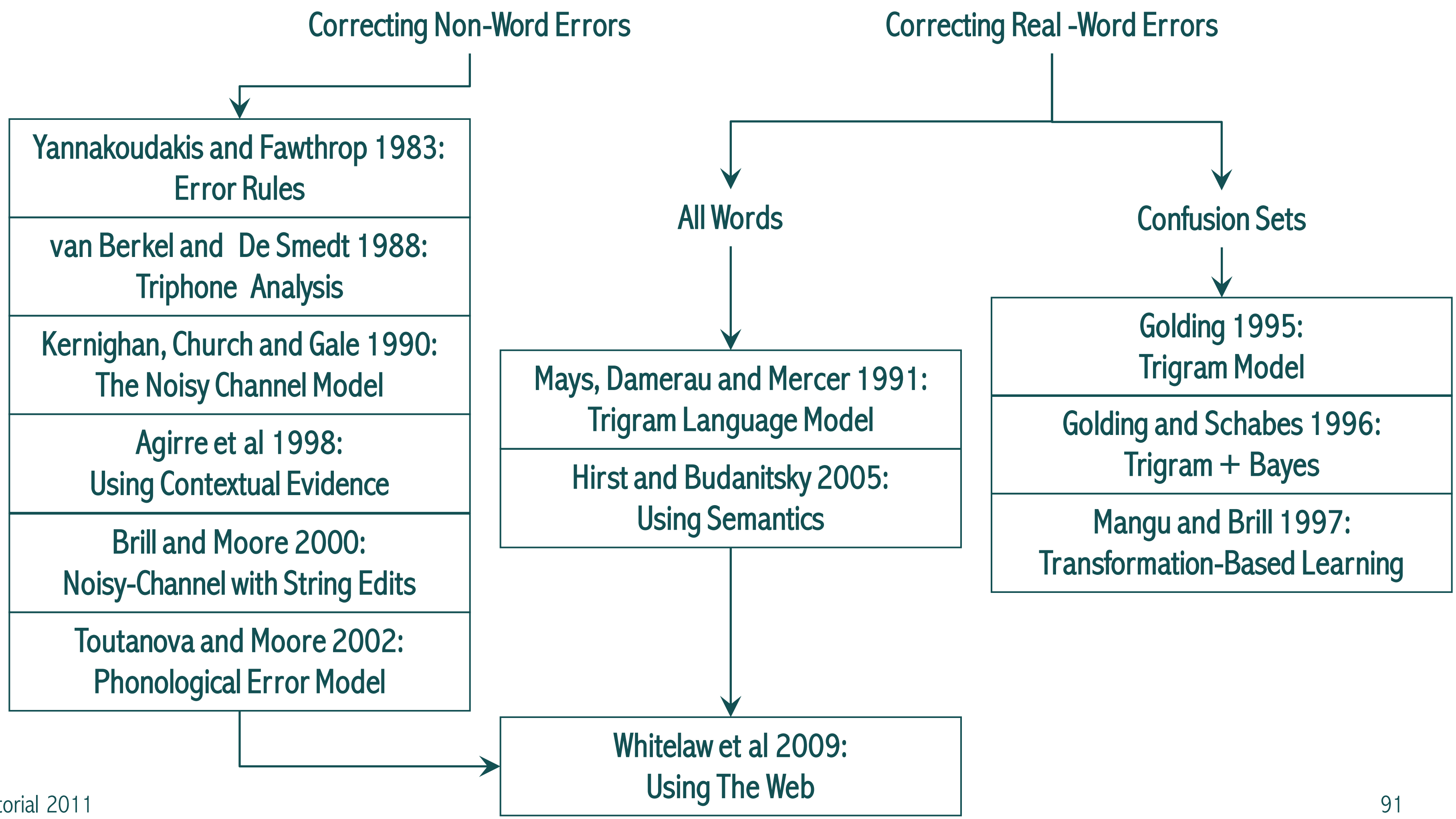
Whitelaw et al [2009]: Candidate Corrections

- The Term List:
  - The 10 million most frequently occurring tokens from a > 1 billion sample of web pages (so it's noisy)
- The Error Model:
  - A substring model like Brill and Moore's
  - Built using ⟨intended word, misspelling⟩ pairs inferred from the web
- The Language Model:
  - Derived from the web, of course

Whitelaw et al [2009]: Performance

- Total error rate for best configuration reduces the error of the best aspell system from 4.83% to 2.62% on artificial data
- Total error rate reduces the error of the best aspell system from 4.58% to 3.80% on human English data
- Total error rate reduces the error of the best aspell system from 14.09% to 9.80% on human German data

A Road Map



Correcting Non-Word Errors
Correcting Real -Word Errors
Yannakoudakis and Fawthrop 1983:
Error Rules
van Berkel and De Smedt 1988:
Triphone Analysis
Kernighan, Church and Gale 1990:
The Noisy Channel Model
Agirre et al 1998:
Using Contextual Evidence
Brill and Moore 2000:
Noisy-Channel with String Edits
Toutanova and Moore 2002:
Phonological Error Model
All Words
Mays, Damerau and Mercer 1991:
Trigram Language Model
Hirst and Budanitsky 2005:
Using Semantics
Whitelaw et al 2009:
Using The Web
Confusion Sets
Golding 1995:
Trigram Model
Golding and Schabes 1996:
Trigram + Bayes
Mangu and Brill 1997:
Transformation-Based Learning

Conclusions

- Methods for generating candidate corrections for a word known to be in error are now very sophisticated
  - The noisy channel model is a good fit
  - Lots of scope for refinement in the language model and the error model
- Determining when a word has been misspelled as another word is an AI-hard problem …
- … but Google-scale language modelling does surprisingly well

Overview

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

The Need

Products » WhiteSmoke Writer 2011 » Grammar Checker

Grammar Checker

You might have asked yourself before sending an important email to a business colleague or a new friend:

> *"Will this text read better if I perform a grammar check?"*

You are not alone! People all around the world find themselves asking this question when trying to avoid grammar mistakes in their texts.

Grammar Checker – The Ultimate Solution for Your Grammar Errors

Proofread your text in a single click by using an online grammar checker.

An online grammar checker will save you the embarrassment of sending a text with grammar mistakes and will make your text look more professional and reliable.

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM's EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

What is a Grammatical Error?

- Something that breaks the rules of the language
- Who decides?
  - Dialects
  - Formality
  - Language change
- Some jurisdictions are stricter than others
  - L'Académie française and its 40 ‘immortals’

Agreement Errors: The Paradigm Grammatical Error

- John and Mary is coming today.
- The block are red.
- A blocks are red.

Taxonomies of Error: Douglas and Dale 1991

- Spelling Errors
- Syntactic Errors
  - Co-occurrence Errors
    - Number Disagreement
    - Bad Subcategorisation
    - Resumptive Pronoun
    - Syntactic Parallelism
  - Dependency and Subordination Errors
    - Bad Clause Conjunction
    - Misleading PP Attachment
    - Misleading Adverbial Attachment
    - Missing Subordination Indicator
    - Redundancy
- Semantic Problems
- Stylistic Problems
- Rhetorical Problems
- Punctuation Problems

Subject–Verb Number Disagreement

- But the males in this study experienced significant difficulties in this area and this problem suggest that some more attention be paid to the phenomenon.
- This method requires a user to think aloud while performing a task, while the researchers makes notes, and perhaps records the session on audio or video tape.
- The main reported problems was the Unix editor vi.

Subject–Verb Number Disagreement

- But the males in this study experienced significant difficulties in this area and this problem suggest that some more attention be paid to the phenomenon.
- This method requires a user to think aloud while performing a task, while the researchers makes notes, and perhaps records the session on audio or video tape.
- The main reported problems was the Unix editor vi.

→The main reported problems were with the Unix editor vi.

Incorrect Subcategorisation Frames: Verbs

- Both Carroll's work and our own, however, has tended to use existing commercial manuals as a basis --- and the question then is how to prune to a fraction of their original size, and to alter their contents to approach more closely to the problems that users actually confront when trying to learn a new system.

Incorrect Subcategorisation Frames: Verbs

- Both Carroll's work and our own, however, has tended to use existing commercial manuals as a basis --- and the question then is how to prune to a fraction of their original size, and to alter their contents to approach more closely ~~to~~ the problems that users actually confront when trying to learn a new system.

Incorrect Subcategorisation Frames: Nouns and Prepositions

- Their feedback pointed to problem areas and causes for misinterpretation, and suggestions of improvements offered by them.

Incorrect Subcategorisation Frames: Nouns and Prepositions

- Their feedback pointed to problem areas and causes <u>for</u> misinterpretation, and suggestions <u>of</u> improvements offered by them.

→Their feedback pointed to problem areas and causes of misinterpretation, and suggestions for improvements offered by them.

Incorrect Subcategorisation Frames: Verbs

- In this way, it is anticipated that the issue of native users not really knowing what it is they need to know is dealt with.

Incorrect Subcategorisation Frames: Verbs

- In this way, <u>it is anticipated that</u> the issue of native users not really knowing what it is they need to know <u>is dealt with</u>.

→In this way, it is anticipated that the issue of native users not really knowing what it is they need to know will be dealt with.

Incorrect Subcategorisation Frames: Nouns and Prepositions

- All mailing systems have capabilities of composing, sending and receiving messages.

Incorrect Subcategorisation Frames: Nouns and Prepositions

- All mailing systems have capabilities of composing, sending and receiving messages.

→All mailing systems have facilities for composing, sending and receiving messages.

Incorrect Subcategorisation Frames: Adjectival Complements

- The feature checklist was easy to administer and complete by experienced users …

Incorrect Subcategorisation Frames: Adjectival Complements

- The feature checklist was easy to administer and complete by experienced users …

→The feature checklist was easy to administer and easy for experienced users to complete …

Syntactic Parallelism Failures

- Semi-structured interviews were conducted with experienced users to find what their most common tasks, the tasks a new user would need to begin, and what errors would be most likely in the early stages.

Syntactic Parallelism Failures

- Semi-structured interviews were conducted with experienced users to find what their most common tasks, the tasks a new user would need to begin, and what errors would be most likely in the early stages.

→Semi-structured interviews were conducted with experienced users to find what their most common tasks were, what tasks a new user would need to begin, and what errors would be most likely in the early stages.

Bad Clause Conjunction

- It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and significantly better performance on individual subtasks including recovery from error.

Bad Clause Conjunction

- It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and significantly better performance on individual subtasks including recovery from error.

→It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and it gave significantly better performance on individual subtasks including recovery from error.

Bad Clause Conjunction

- The conditions under which our subjects worked tended to minimize such problems – since we asked them to persevere, and in the end they would be able to get human help.

Bad Clause Conjunction

- The conditions under which our subjects worked tended to minimize such problems – since we asked them to persevere, and in the end they would be able to get human help.

→The conditions under which our subjects worked tended to minimize such problems, since we asked them to persevere, and in the end they would be able to get human help.

Bad Clause Conjunction

- The more active but ineffectual behaviour of the males may mean that they feel they must be capable of mastering the system, of overcoming their errors and are less worried or affected by the possibility of making errors.

Bad Clause Conjunction

- The more active but ineffectual behaviour of the males may mean that they feel they must be capable of mastering the system, of overcoming their errors and are less worried or affected by the possibility of making errors.

→The more active but ineffectual behaviour of the males may mean that they feel they must be capable of mastering the system and of overcoming their errors, and are less worried or affected by the possibility of making errors.

Bad Clause Conjunction

- Novice users should, however, be able to voice thoughts and desires on any topic, throughout the process if the manual is to be properly user-centred.

Bad Clause Conjunction

- Novice users should, however, be able to voice thoughts and desires on any topic, throughout the process if the manual is to be properly user-centred.

→However, if the manual is to be properly user-centred, novice users should be able to voice thoughts and desires on any topic throughout the process.

Syntactic Redundancy

- So although this seems to be is a winning feature in learning, it may not …
- … this problem suggests that some more attention be paid to the phenomenon
- … thus so this argues for the complementary use of …

Syntactic Redundancy

- So although this seems to be is a winning feature in learning, it may not …
- … this problem suggests that some more attention be paid to the phenomenon
- … thus so this argues for the complementary use of …

What Causes Grammar Errors?

- Competence-based errors (Errors of Intention):
  - Unfamiliarity with the language
- Performance-based errors (Errors of Execution):
  - Repeated words
  - Editing errors

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM’s EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

The Unix Writer’s Workbench

- A breakthrough in the early 1980s
  - We believe that the Writer's Workbench programs provide a more general text analysis system than JOURNALISM or CRES, and unlike EPISTLE they are already in wide use. At Bell Laboratories there are over 1000 users on over 50 machines. [1982:106]
- Widely-used in educational contexts
- Underlying technology formed the basis for the first PC grammar checkers: Grammatik, RightWriter, StyleWriter

The Unix Writer's Workbench: Proofreading with PROOFR

- Checks for existence of non-word spelling errors; user-specified automatic correction can be carried out
- Checks for unbalanced punctuation and other simple punctuation mistakes
- Checks for double words
- Checks for misused words, wordy phrases, sexist terms, …
- Checks for split infinitives using a simple PoS tagger

The Unix Writer's Workbench: Stylistic Analysis with STYLE

- Based on PoS tagging, provides 71 numbers describing stylistic features of the text
  - Readability indices
  - Average sentence and word length
  - Distribution of sentence lengths
  - Percentage of verbs in passive voice
  - Percentage of nouns that are nominalisations
  - …

The Unix Writer's Workbench: Stylistic Analysis with STYLE

```
readability grades:
        (Kincaid) 11.3  (auto) 12.6  (Coleman-Liau) 13.1  (Flesch) 13.2 (48.8)
sentence info:
        no. sent 240 no. wds 4636
        av sent leng 19.3 av word leng 5.18
        no. questions 1 no. imperatives 0
        no. content wds 2734  59.0%   av leng 6.72
        short sent (<14) 24% (58) long sent (>29)   9% (22)
        longest sent 64 wds at sent 150; shortest sent 4 wds at sent 70
sentence types:
        simple  42% (101) complex  38% (92)
        compound   7% (16) compound-complex  13% (31)
word usage:
        verb types as % of total verbs
        tobe  32% (170) aux  16% (85) inf  17% (89)
        passives as % of non-inf verbs  14% (63)
        types as % of total
        prep 10.5% (487) conj 3.8% (177) adv 4.2% (197)
        noun 28.0% (1296) adj 17.2% (797) pron 4.7% (220)
        nominalizations   2 % (90)
sentence beginnings:
        subject opener: noun (48) pron (28) pos (1) adj (35) art (57) tot  70%
        prep  13% (32) adv   6% (15)
        verb   1% (3)  sub_conj   6% (14) conj   2% (5)
        expletives   1% (2)
```

The Unix Writer's Workbench: Other Components

- PROSE: compares the stylistic parameters of a given text against a domain-specific standard
- ABST: determines the conceptual abstractness of a text via a list of 314 abstract words
- ORG: prints only first and last sentences of paragraphs

Atwell [1987]: CLAWS

- Originally built to assign PoS tags to the London-Oslo-Bergen corpus
- Developed in part because of the computational cost of more complex systems:
  - ‘[Heidorn et al 82] reported that the EPISTLE system required a 4Mb virtual machine (although a more efficient implementation under development should require less memory).’ [1987:38]

Atwell [1987]: Constituent-Likelihood Error Detection

- For PoS tagging, uses a table of PoS bigram frequencies to determine most likely sequences
- Detects grammatical errors by flagging unlikely PoS transitions
- Doesn't need separate data for training error likelihoods

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM's EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

IBM's EPISTLE: History

- Initial work in the early 1980s led to several innovative techniques
- Based on Heidorn's Augmented Phrase Structure Grammar [1975]
- Renamed CRITIQUE somewhere in the mid to late1980s
- Released on IBM mainframes late 1980s
- Key team members went on to build Microsoft Word's grammar checker from 1992 onwards
- Grammar checking released as part of MS Word 97

IBM's CRITIQUE: Grammar vs Style

- Grammatical critiques:
  - Strict rules as to whether a sentence is grammatical or not
  - Correction is typically clear
- Stylistic weaknesses are less black and white:
  - too great a distance between subject and verb
  - too much embedding
  - unbalanced subject/predicate size
  - excessive negation or quantification
  - …

IBM's CRITIQUE : Grammar Errors

- Number Disagreement:
  - he go, many book, it clarifies and enforce
- Wrong Pronoun Case:
  - between you and I, it is me
- Wrong Verb Form:
  - had expect, seems to been
- Punctuation:
  - run-on sentences, questions with a final period instead of a question mark
- Confusions:
  - who's vs whose, it's vs its, your vs you're, form vs from

IBM's CRITIQUE : Stylistic Weaknesses #1

- Excessive length
  - Sentences or lists that are too long
  - Sequences with too many prepositional phrases
- Excessive complexity
  - Noun phrases with too many premodifiers
  - Clauses with a series of *and*s
  - Verb phrases with too many auxiliary verbs
  - Clauses with too much negation
- Lack of parallelism
  - Example: you should drink coffee rather than drinking tea

IBM's CRITIQUE : Stylistic Weaknesses #2

- Excessive formality
  - phrases that are bureaucratic, pompous or too formal
- Excessive informality
  - constructions acceptable in spoken English but too informal when written
- Redundancy
  - phrases that can be shortened without loss in meaning
- Missing punctuation
- Nonpreferred constructions
  - Split infinitives [eg to completely remove], colloquial usage [eg ain't working]

The MS Word Grammar Checker: Processing Steps

1. Tokenisation and Lexical Lookup
2. Syntactic Sketch
3. Syntactic Portrait
4. Production of Logical Forms

The MS Word Grammar Checker: An Example

- Consider the following sentence:
  - After running a mile he seemed tired.

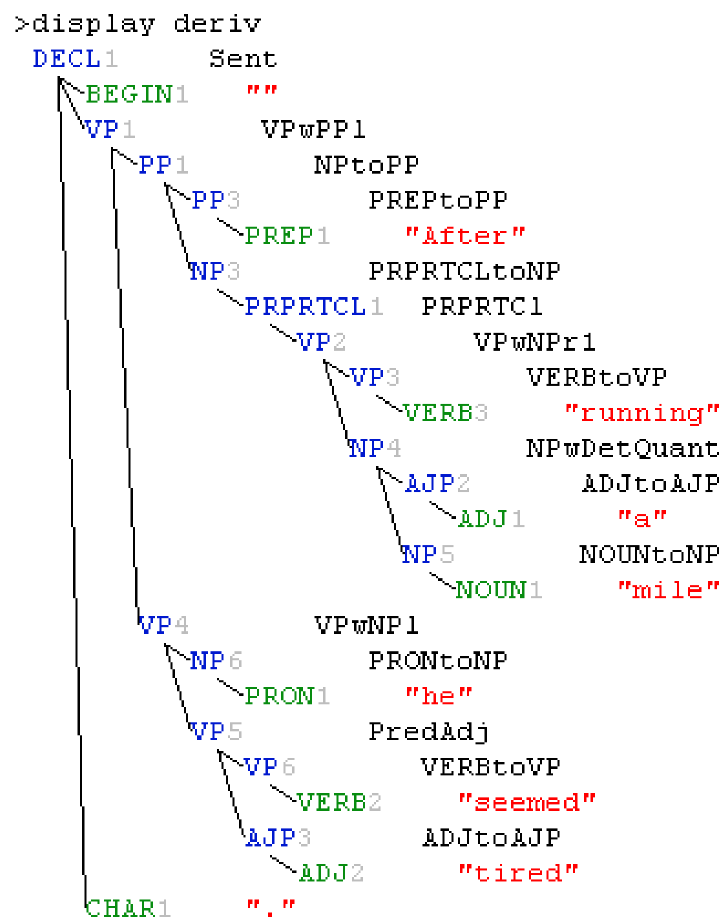
The MS Word Grammar Checker: Lexical PoS Records

- Also includes detection of multiword elements and named entity mentions
- Lexicon based on LDOCE and AHD + supplementary information added both manually and automatically
- Over 100k words
- There are two other records produced for 'after' here for the Adj and Adv uses

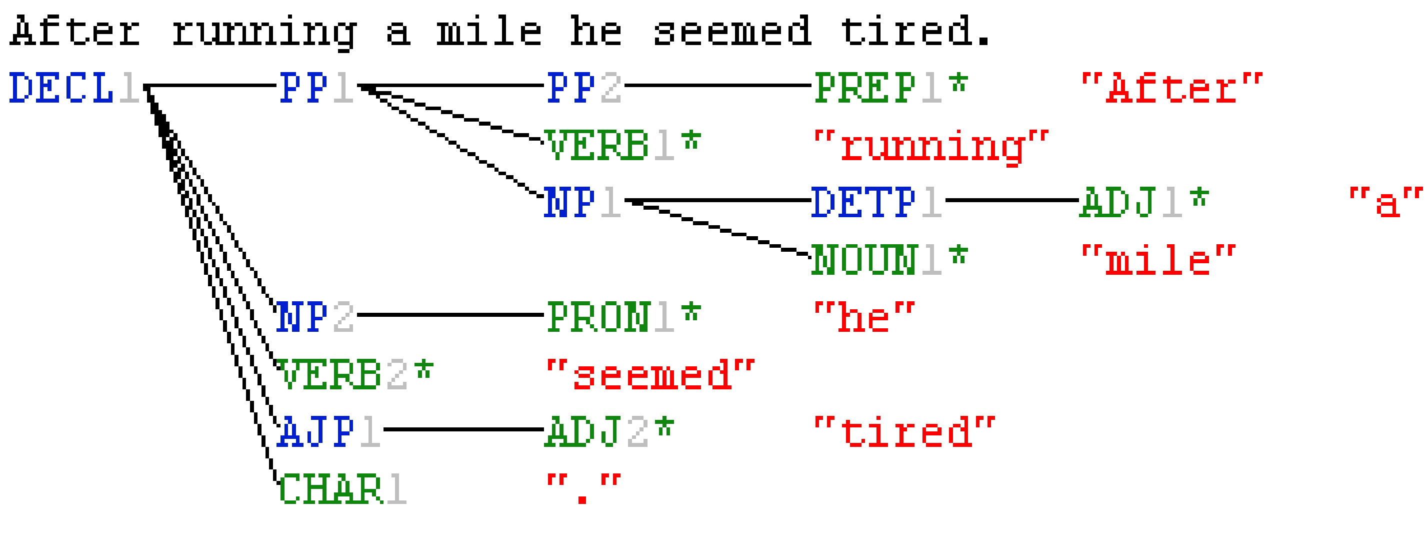
```
 ------------------------------
{Segtype     PREP
 Nodetype    PREP
 Nodename    PREP1
 Ft-Lt       1-1
 String      "After"
 CopyOf      REC40
 Lex         "After"
 Lemma       "after"
 Bits        TakesAn InitCap Tme
 Prob        1.00000 }
 ------------------------------
{Segtype     CONJ
 Nodetype    CONJ
 Nodename    CONJ1
 Ft-Lt       1-1
 String      "After"
 CopyOf      REC41
 Lex         "After"
 Lemma       "after"
 Bits        Subconj TakesAn
             InitCap Tme
 Prob        0.00119 }
 ------------------------------
```

The MS Word Grammar Checker: Syntactic Analysis

- Bottom-up chart parser
- Uses probabilities and heuristics
- Grammar contains 125 mostly binary rules
- This is the <u>derivation tree</u>



The MS Word Grammar Checker: Syntactic Analysis



The MS Word Grammar Checker: Syntactic Information Stored at the Root Node

```
>display record DECL1
{Segtype     SENT
 Nodetype    DECL
 Nodename    DECL1
 Ft-Lt       0-8
 String      " After running a mile he seemed tired ."
 CopyOf      VP1
 Rules       (Sent VPwPP1 VPwNP1 PredAdj VERBtoVP)
 Constits    (BEGIN1 VP1 CHAR1)
 Lex         "seemed"
 Lemma       "seem"
 Bits        Pers3 Sing Past Closed
             L9 B0 Wv6 L1 L7 I3
             RtoSub F0 Wv8 Wv7 Wv4
             Wv6N I5 I6
 Prob        0.25645
 Prmods      PP1 "After running a mile"
             NP2 "he"
 Head        VERB2 "seemed"
 Psmods      AJP1 "tired"
             CHAR1 "."
 Subject     NP2 "he"
 Predadj     AJP1 "tired"
 Props       DECL1 " After running a mile he seemed tired ."
             PRPRTCL1 "running a mile"
 Pod         40
 Inverts     PP1 "After running a mile"
 Nargs       1
 FrstV       VERB2 "seemed"
 Vprp        (like)
 Predicat    VP5 "seemed tired"
 Topic       NP2 "he"
 TopPPs      PP1 "After running a mile"
 Score       40.000000000 }
```

The MS Word Grammar Checker: The VP→ VP PP Rule [Abbreviated]

```
VPwPPl:

   PP ( ^Comma(Prp) & ^Nappcomma(lastrec) & ^Precomma(lastrec) & ^SuspSUBCL & (forany(Prmods, [Comma]) -> Coords) &
        forall(firstrecs(PPobj), [Digits^=3 & Digits^=4]) & (forany(lastrecs, [Comma & ^Paren]) -> (Multcomma | Comma(lastphr))) &
        forall(lastrecs, [^Nomcomp | ^T5 | (Compl & Lemma(lasttokn)^="that")]) & (Gerund -> (^Rel(Postadv) | Postadv^=lastrec)) &
        Lemma(Prp) ^in? set{a an but x X} & forall(Coords, [Lemma(Prp) ^in? set{a an but x X}]) )

   VP ( ^Semiaux & ^Relpn & ^Paren &
        (forany(lastrecs(PP), [Nappcomma]) -> (^Pastpart | ^PPobj(first(Psmods)) |
           ^Comma(first(Psmods)))) &
        forall(lastrecs(PP), [Nappcomma -> (^Multcomma | Numbr ^agree? Numbr(VP))]) &
        (Nodetype(lastrec(PP))=="RELCL" -> (^Thatcomp(lasttokn(PP)) |
           Rel(first(Prmods(lastrec(PP)))))) & Nodetype(last(Psmods)) ^in? set{SREL TAG} &
        (Ord(Adj(Lex(lasttokn(PP)))) -> ^Num(Adj(Lex(firsttokn(first(Prmods)))))) &
        (Adv(Lex) -> (Prmods | Obj1 | (^Confus & Lemma ^in? set{no yes}))) & (Wh(Conj(Lex(PP))) -> (Prmods(PPobj(PP)) | YNQ)) &
        (Digits(first(Prmods)) -> (^Comma(first(Prmods)) | Prmods(first(Prmods)) | Nodetype(lasttokn(PP))^="NOUN")) &
        (Mnth(lasttokn(PP)) -> (^Ord(firsttokn) | ^Digits(firsttokn) | Digits(firsttokn)>2)) &
        ((Nom(Pron(Lex(lastrec(PP)))) & ^Obj(Pron(Lex(lastrec(PP))))) ->
           (Subject & Subject in? Prmods)) &  (T5 -> (^Comma | (forall(Psmods, [^Oldsubcl]) &
           (^Nomcomp(Predcomp) | Compl(Predcomp) | ^Comma(lastphr(PP)))))) )

 --> VP { Prmods=PP++Prmods; Props=Props(PP)++Props; -SuspNREL;
         if (Subject(VP) ^in? Prmods(VP) & FortoPP(PP)) {Subject=PP; -VPInvert;}
         else if ((^Subject(VP) | VPInvert(VP)) & ^theresubj_test(VP)) MidPPs=PP++MidPPs;
         else {TopPPs=PP++TopPPs; Inverts=PP++Inverts;}; Pod=Pod+Pod(PP);
         if (Lemma(lasttokn(PP))==";") Pod=Pod-4;
         if (^PPobj(PP) & Loc(Adv(Lex(PP)))) Pod=Pod-1;
         if (Subject in? Prmods(VP) | theresubj_test) Pod=Pod+1; }
```

The MS Word Grammar Checker: A Logical Form

```
seem1 (+Past +L7)
  Dsub-----he1 (+Masc +Pers3 +Sing +FindRef +Anim +Humn)
  Dadj-----tired1 (+F0 +Psych)
  after----run1 (+T1 +Middle +Mov +Loc_sr)
             Dsub-----he1
             Dobj-----mile1 (+Indef +Pers3 +Sing +Conc +Count +Dst)
```

The MS Word Grammar Checker: An Error Checking Rule

```
Desc_Comma5:

  SYNREC (((Nodetype in? set{SUBCL AVP PRPRTCL AVPNP INFCL}) |
               (Nodetype=="PP" & PPobj)) &
           seg==first(Prmods(Parent)) &
           Nodetype(lasttokn) ^= "CHAR" &
           ^Theresubj &
           seg ^= Subject(Parent) &
           (Nodetype=="AVP" -> (^TheAVP & ^forany(Prmods,[TheAVP]))) &
           (Wh -> Lemma=="however") &
           ^forany(Coords,[Wh]) &
           (Nodetype(Head(Parent))=="VERB" | VPcoord(Parent)) &
           (Neg -> ^YNQ(Parent)) &
           ((Subject(Parent) &
              ((Ft(Subject(Parent))<Ft(FrstV(Parent)) & Ft(Subject(Parent))>Ft) |
               (VPcoord(Parent) & Ft(Subject)<Ft(FrstV(first(Coords(Parent)))))) |
             Nodetype(Parent)=="IMPR" |
             (Nodetype(Parent)=="QUES" & (YNQ(Parent) | WhQ(Parent)))))

 --> SYNREC { { segrec rec, commarec;
               commarec=segrec{Nodetype="CHAR"; Lemma=",";};
               rec=segrec{%%SYNREC; Psmods=Psmods++commarec;};
               add_descrip("Comma with Adverbials",0,rec); }; }
```

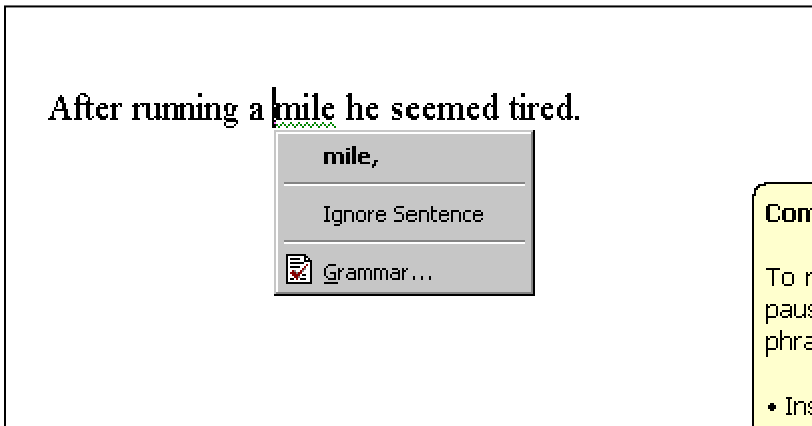
The MS Word Grammar Checker: A Segment Record with An Error

```
>display record PP1
{Segtype     PP
 Nodetype    PP
 Nodename    PP1
 Ft-Lt       1-4
 String      "After running a mile"
 CopyOf      NP3
 Rules       (TrLF_ControlatVP Desc_Comma5 NPtoPP PRPRTCLtoNP PRPRTCl VPwNPr1 VERBtoVP)
 Constits    (PP1 PP1 PP3 NP3)
 Lex         "running"
 Lemma       "run"
 Bits        Pers3 Sing L9 X9 Wv6
             IO D1 T1 L1 L7 T5
             Asubj Loc_sr Unacc Mov
             Middle Wv4
 Prob        0.05383
 Prmods------PP2 "After"
 Head--------VERB1 "running"
 Psmods------NP1 "a mile"
 Gerund------VERB1 "running"
 PPobj-------NP3 "running a mile"
 Prp---------PP2 "After"
 Obj1--------NP1 "a mile"
 Props-------PRPRTCL1 "running a mile"
 Pod         16
 Parent------DECL1 "After running a mile he seemed tired ."
 Nargs       1
 FrstV-------VERB3 "running"
 Object------NP1 "a mile"
 Vptc        (along around away back down in off on out over through up across after)
 Vprp        (across after at from with over into on of through to against)
 Descrips
       {Ft-Lt       1-4
        Value       18
        DescType    "Comma with Adverbials"
        DescRepl----PP4 "After running a mile"
        DescReplStr "After running a mile," }
 SemNode-----run1
 PrevCat     PP }
```

The MS Word Grammar Checker: The Results of Error Checking

```
>display desc

Comma with Adverbials:
        After running a mile     consider:  After running a mile,
```



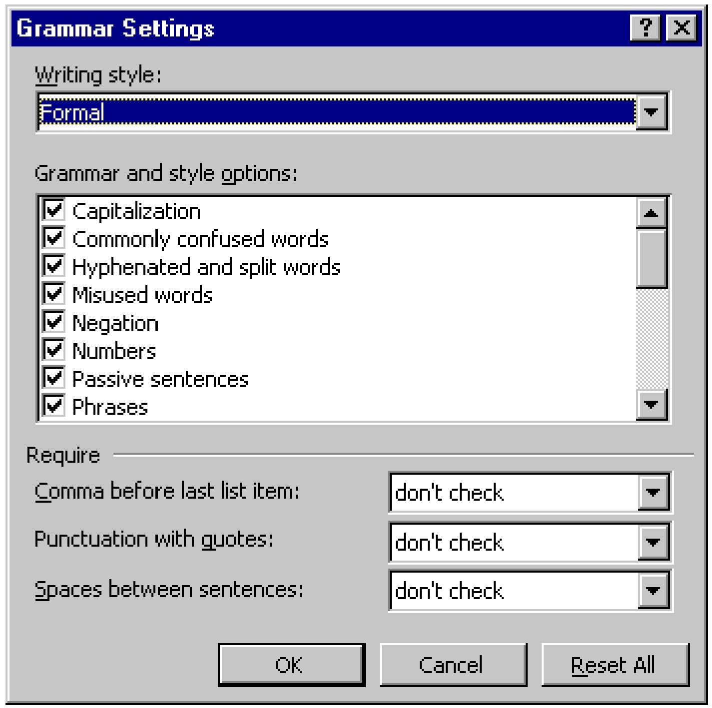
Comma Use

To make your sentence easier to read or to signal a pause, consider using a comma to set off words or phrases (especially introductory words or phrases).

- Instead of: Unfortunately it rained the day of the picnic.
- Consider: Unfortunately, it rained the day of the picnic.

- Instead of: Once he got home he began to calm down.
- Consider: Once he got home, he began to calm down.

The MS Word Grammar Checker: Controlling the Checker's Behaviour



EPISTLE/CRITIQUE/MS Word: Key Ideas

- A metric for ranking alternative parses [Heidorn 1982]
- Relaxation for parsing errorred sentences [Heidorn et al 1982]
- A heuristic fitted parsing technique for sentences outside the grammar's coverage [Jensen et al 1983]

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM's EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

Constraint Relaxation: The Basic Idea

- When a sentence cannot be parsed, <u>relax</u> the grammar rules in some way so that it can be parsed
- The particular constraints that are relaxed indicate what the nature of the grammatical error is
- First explored in the context of robust parsing by Weischedel and Black [1980]

Constraint Relaxation: Handling Constraint Violation Errors

- Subject-verb number agreement
  - * John and Mary runs
- Premodifier-noun number agreement
  - * This dogs runs
- Subject-complement number agreement
  - * There is five dogs here
- Wrong pronoun case
  - * He and me ran to the door
- Wrong indefinite article
  - * A apple and a rotten old pear.

Constraint Relaxation: Handling Constraint Violation Errors

- **A number agreement constraint in PATR-II:**

$$
\begin{array}{llll}
\mathrm{X0} & \rightarrow \mathrm{X1\ X2} & & \\
 & \langle \mathrm{X0\ cat} \rangle & = & \mathrm{VP} \\
 & \langle \mathrm{X1\ cat} \rangle & = & \mathrm{NP} \\
 & \langle \mathrm{X2\ cat} \rangle & = & \mathrm{VP} \\
 & \langle \mathrm{X0\ subject} \rangle & = & \mathrm{X1} \\
 & \langle \mathrm{X1\ num} \rangle & = & \langle \mathrm{X2\ num} \rangle
\end{array}
$$

Constraint Relaxation [Douglas and Dale 1992]: Relaxation Packages

X0 $\rightarrow$ X1 X2

1	$\langle$X0 cat$\rangle$	=	NP
2	$\langle$X1 cat$\rangle$	=	Det
3	$\langle$X2 cat$\rangle$	=	N
4	$\langle$X1 agr precedes$\rangle$	=	$\langle$X2 agr begins$\rangle$
5	$\langle$X1 agr num$\rangle$	=	$\langle$X2 agr num$\rangle$
6	$\langle$X0 agr num$\rangle$	=	$\langle$X2 agr num$\rangle$

Relaxation level 0:
necessary constraints = {1,2,3,4,5,6}
optional constraints = {}

Relaxation level 1:
necessary constraints: {1,2,3}
relaxation packages:

(a) {5, 6}: Premodifier-noun number disagreement
(b) {4}: *a/an* error

Constraint Relaxation

- Advantages:
  - provides a precise and systematic way of specifying the relationship between errorful and ‘correct’ forms, making it easier to generate suggestions for corrections
- Disadvantages:
  - Requires significant amounts of hand-crafted linguistic knowledge

Mal-Rules

- Also known as error anticipation
- Mal-rules explicitly describe specific expected error forms

A Mal-Rule for Handling Omissions [Schneider and McCoy 1998]

- Example:

  The boy happy

- Conventional rule:

  VP → V AdjP

- Malrule:

  VP[error +] → AdjP

Mal-Rules

- Advantage:
  - Specifically targets known problems
  - Allows easy identification of the nature of the error
- Disadvantages:
  - Requires error types to be catalogued in advance
  - Infeasible to anticipate every possible error
- Arguably mal-rules are just a notational variant of constraint relaxation approaches

Other Approaches

- Fitted parsing [Jensen et al 1983]
- Mixed bottom-up and top-down parsing [Mellish 1989]
- Minimum edit distance parsing [Lee et al 1995]

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM's EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

Robust Parsing

- The Goal:
  - Analyse extragrammatical input in order to extract some useful meaning
- No need to characterise and repair the error
- Processing of spoken language is a special case

Controlled Languages

- The Goal:
  - Ensure that a text conforms to a specific set of rules and conventions
- Examples:
  - ASD Simplified Technical English
  - Caterpillar Technical English
  - EasyEnglish
  - Attempto Controlled English
- See http://www.geocities.ws/controlledlanguage/

Outline

- What is a Grammatical Error?
- Grammar Checking without Syntax
- IBM's EPISTLE
- Grammar Checking Techniques
- Related Areas
- Commercial Packages

Do Current Grammar Checkers Help?

- In real use, grammar checkers may have low recall <u>and</u> low precision

Kohut and Gorman [1995]: An Empirical Evaluation of Five Packages

Package	Total # Errors	Real Errors Correctly Identified	Real Errors Incorrectly Identified	False Errors	False Errors/Total Deteted
PowerEdit	133	47%	12%	11%	16.13%
RightWriter	133	34%	8%	7%	13.85%
Grammatik	133	31%	6%	11%	23.44%
Editor	133	17%	3%	4%	16.13%
CorrectGrammar	133	15%	5%	10%	32.5%

Kohut and Gorman [1995]: An Empirical Evaluation of Five Packages

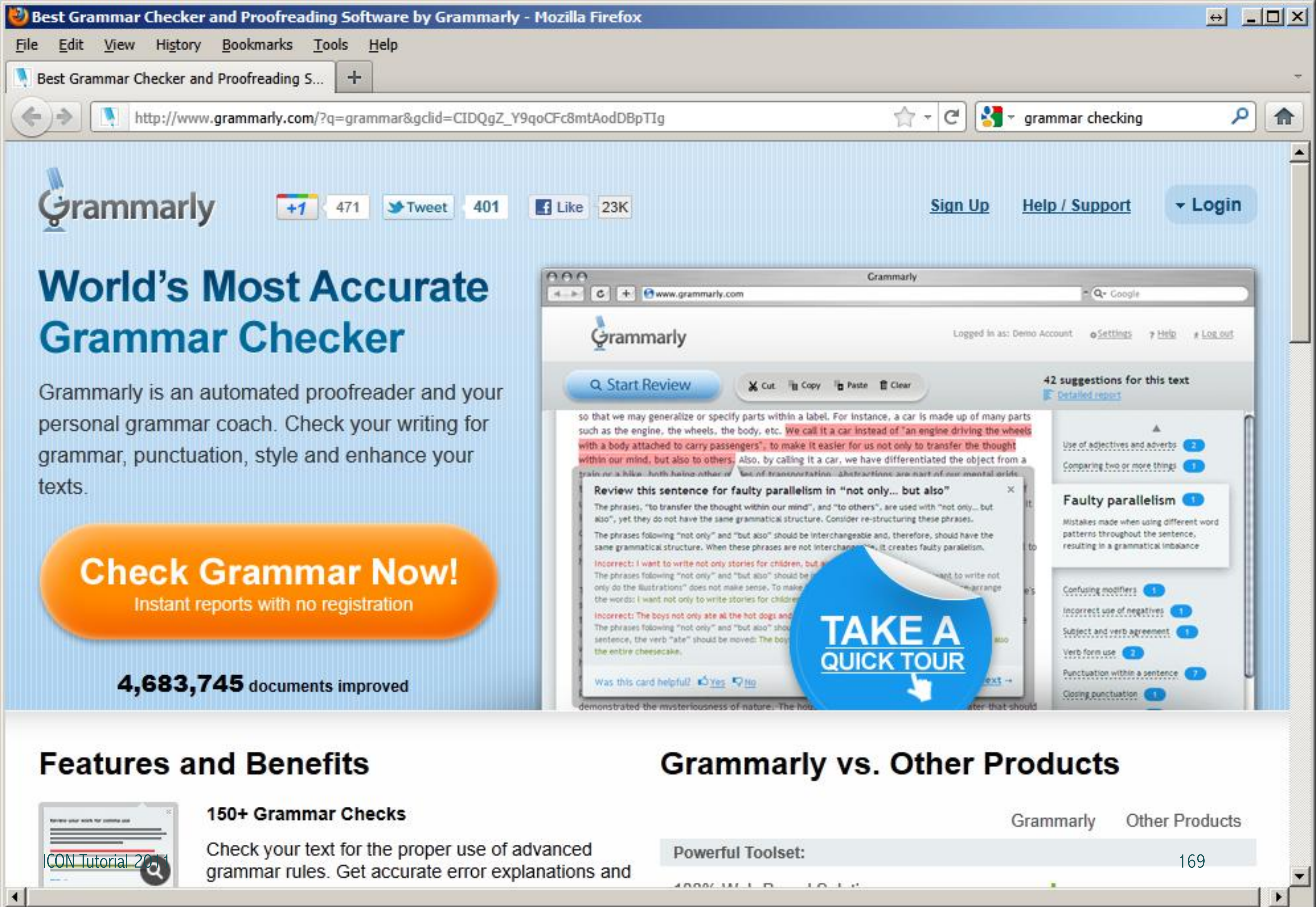
Mechanical Errors

	Errors Found by Authors	PowerEdit		RightWriter		Grammatik		Editor		Percentage Correct Grammar	
Punctuation	29	13	(45%)	9	(31%)	5	(17%)	5	(17%)	3	(10%)
Agreement	8	2	(25%)	2	(25%)	3	(38%)	1	(13%)	2	(25%)
Capitalization	2	0	(0%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Verb form	3	1	(33%)	1	(33%)	3	(100%)	0	(0%)	0	(0%)
Sentence structure	20	15	(75%)	10	(50%)	9	(45%)	2	(10%)	3	(15%)
Total mechanical errors	62	31	(50%)	22	(35%)	20	(32%)	8	(13%)	8	(13%)

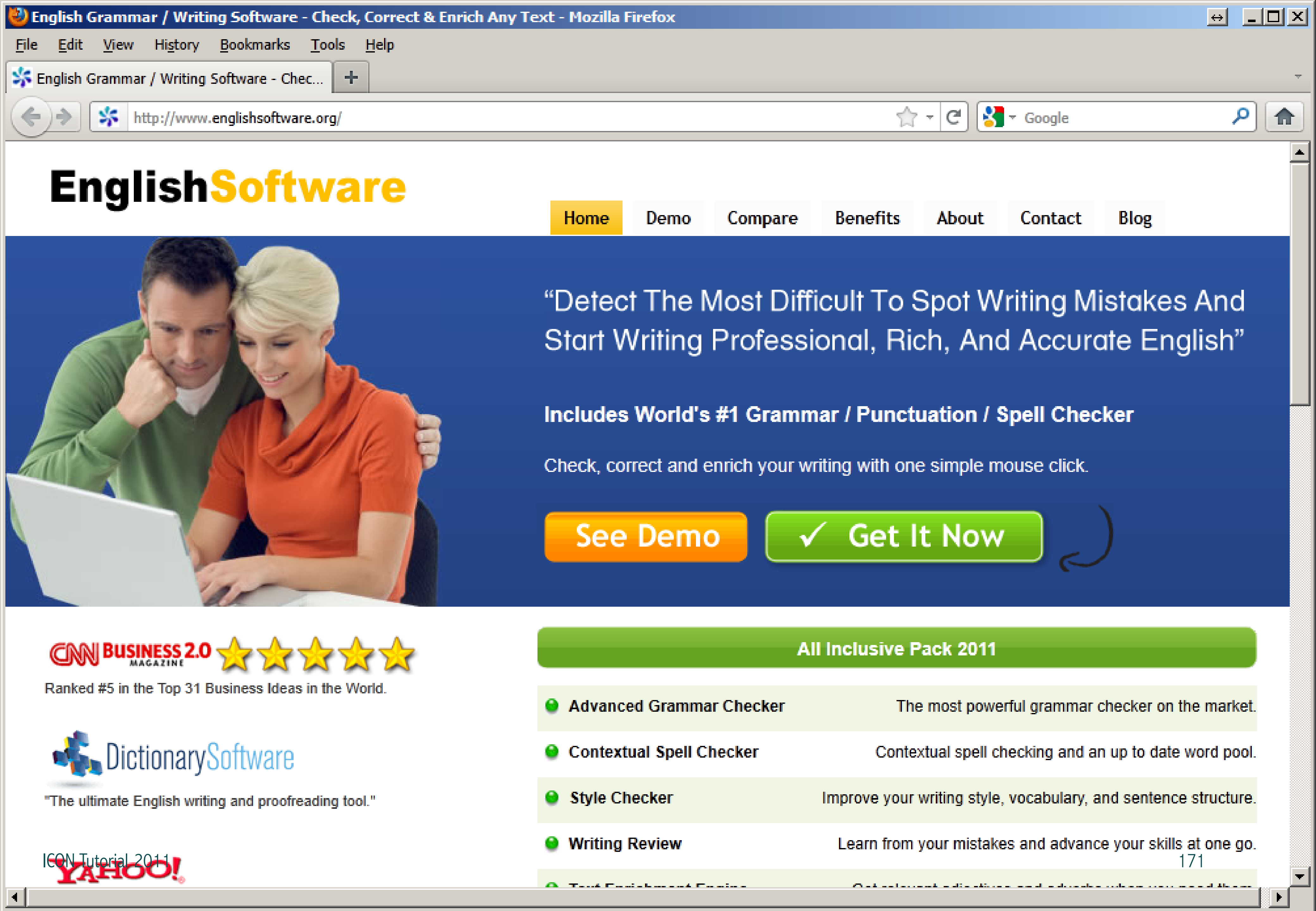
Style Errors

	Errors Found by Author	PowerEdit		RightWriter		Grammatik		Editor		Correct Grammar	
Passive voice	15	9	(60%)	7	(47%)	4	(27%)	0	(0%)	7	(47%)
Complex sentences	3	3	(100%)	3	(100%)	2	(67%)	0	(0%)	1	(33%)
Wrong word	21	3	(14%)	4	(19%)	7	(33%)	4	(19%)	3	(14%)
Redundancy	5	1	(20%)	0	(0%)	0	(0%)	1	(20%)	0	(0%)
Weak wording	18	13	(72%)	7	(39%)	5	(28%)	6	(33%)	1	(6%)
Slang/colloquialisms	2	1	(50%)	1	(50%)	0	(0%)	1	(50%)	0	(0%)
Sexist language	6	0	(0%)	1	(17%)	1	(17%)	2	(33%)	0	(0%)
Negative wording	1	1	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
Total style errors	71	31	(44%)	23	(32%)	19	(27%)	14	(20%)	12	(17%)


StyleWriter Software (Style Writer Software) - Your Plain English Manuscript Editing Software, Spell Checker, and English Grammar Checker Software - Mozilla F...
File Edit View History Bookmarks Tools Help
StyleWriter Software (Style Writer Software...
http://www.stylewriter-usa.com/
stylewriter
Free Download StyleWriter software
World's Largest Writing Style, Word Usage, Readability and Grammar Checker developed by Editor Software
YOUR Plain English Manuscript Editing Software
StyleWriter 4 version
http://www.StyleWriter-USA.com
Home | Contact | Free Articles
Home
What is StyleWriter?
Features of StyleWriter
Videos / Demos
Free Download of StyleWriter
StyleEase Products
1. Welcome to StyleWriter
2. What is StyleWriter?
3. How does StyleWriter work?
4. Clarity and Readability
5. StyleWriter's Editing Features
6. StyleWriter's Scoring
StyleWriter 4
The world's most powerful plain English editing software just got better
With powerful new features...
Our newest release, StyleWriter v4.0, is a blazing-fast 32-bit application that is compatible with WINDOWS 7 and MS-WORD 2010! StyleWriter runs on YOUR computer and NEVER needs
http://www.stylewriter-usa.com/


Best Grammar Checker and Proofreading Software by Grammarly - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Best Grammar Checker and Proofreading S...
http://www.grammarly.com/?q=grammar&gclid=CIDQgZ_Y9qoCFc8mtAodDBpTIg
grammar checking
Grammarly
+1 471
Tweet 401
Like 23K
Sign Up
Help / Support
Login
World's Most Accurate Grammar Checker
Grammarly is an automated proofreader and your personal grammar coach. Check your writing for grammar, punctuation, style and enhance your texts.
Check Grammar Now!
Instant reports with no registration
4,683,745 documents improved
Start Review
42 suggestions for this text
Review this sentence for faulty parallelism in "not only... but also"
Faulty parallelism
Mistakes made when using different word patterns throughout the sentence, resulting in a grammatical imbalance
TAKE A QUICK TOUR
Features and Benefits
150+ Grammar Checks
Check your text for the proper use of advanced grammar rules. Get accurate error explanations and
Grammarly vs. Other Products
Grammarly
Other Products
Powerful Toolset:


English Grammar Checker Software | WhiteSmoke | World-Leading Language Solutions by WhiteSmoke - Mozilla Firefox
File Edit View History Bookmarks Tools Help
English Grammar Checker Software | White...
http://www.whitesmoke.com/
whitesmoke
Change Language
WhiteSmoke™
World-Leading English Writing Software
Chat – Talk to a Representative
About
Contact
Support
Home
Products
Store
Blog
Partners
Demo
FAQ
Search
World-Leading English Writing Software
Perfect your English writing with the most advanced editing tools:
Spelling, Grammar Checker, Punctuation, Style and Structure
Full Text Translation & Multilingual Dictionary:
Translate full text in a single click and get usage examples for every word.
Learn from your mistakes with our error explanation tool:
Check every aspect of your text and learn how to avoid the same mistakes in the future.
Get it Now!
Click here
"WhiteSmoke wins the
CNN
"One of the Top 31 Business
"An all-in-one
http://www.whitesmoke.com/partners.html
Get Our Special Offer!


English Grammar / Writing Software - Check, Correct & Enrich Any Text - Mozilla Firefox
File Edit View History Bookmarks Tools Help
English Grammar / Writing Software - Chec...
http://www.englishsoftware.org/
Google
EnglishSoftware
Home
Demo
Compare
Benefits
About
Contact
Blog
"Detect The Most Difficult To Spot Writing Mistakes And Start Writing Professional, Rich, And Accurate English"
Includes World's #1 Grammar / Punctuation / Spell Checker
Check, correct and enrich your writing with one simple mouse click.
See Demo
✓ Get It Now
CNN BUSINESS 2.0 MAGAZINE
Ranked #5 in the Top 31 Business Ideas in the World.
DictionarySoftware
"The ultimate English writing and proofreading tool."
YAHOO!
All Inclusive Pack 2011
Advanced Grammar Checker
The most powerful grammar checker on the market.
Contextual Spell Checker
Contextual spell checking and an up to date word pool.
Style Checker
Improve your writing style, vocabulary, and sentence structure.
Writing Review
Learn from your mistakes and advance your skills at one go.


Ginger Software - The World's Leading Grammar and Spell Checker - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Ginger Software - The World's Leading Gra...
http://www.gingersoftware.com/
Google
Ginger
Store
Download
Blog
Affiliates
Help
About Us
Media
Educators
The World's #1 Grammar & Spell checker
Free Download
Proofreading, spelling and grammar check in any application
Improve your English writing - learn from your mistakes
Text to speech reader - improve your pronunciation
Compatible with:
Try it yourself
Type/paste your text for correction and Ginger It!
Ginger It!


Paper Rater: Free Online Grammar Checker, Proofreader, and More - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Paper Rater: Free Online Grammar Checker...
http://www.paperrater.com/
Google
+1
Tweet
Like 179
PaperRater
Home
About
Features
Pricing
Help / FAQs
Grammar & Spelling Check
Free Online Proofreading
No Downloads
Use Now FREE!
feedback
Grammar Checking
an
Wrong article
a scorching
Explain...
Ignore suggestion
Ignore all
Robust grammar checking allows you to find those pesky mistakes and correct them before turning in your paper.
Plagiarism Detection
Originality / Plagia
Original Work
Originality: 100%
No sign of plagiar
seen!
Find out if your paper contains plagiarized text before your professor does.
Writing Suggestions
Word Choice
Style
Vocabulary Words
Our proofreading system alerts you to opportunities to improve your writing.

Paper Rater - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Paper Rater

http://www.paperrater.com/submit_paper

paper rater

PaperRater

Title Some Sample Grammar Problems

John and Mary is coming today.

A blocks are

Wrong Auxiliary Verb

- blocks are
- A block are
- Explain...
- Ignore suggestion
- Ignore all
- Edit Selection...

But the male ... ed significant difficulties in this area and this problem suggest that some more attention be ... n.

This method ... aloud while performing a task, while the researchers makes notes, and perhaps records the ... tape.

The main re ... Unix editor vi.

Both Carroll' ... vever, has tended to use existing commercial manuals as a basis --- and the question the ... action of their original size, and to alter their contents to approach more closely to the probl ... onfront when trying to learn a new system.

Their feedback pointed to problem areas and causes for misinterpretation, and suggestions of improvements offered by them.

In this way, it is anticipated that the issue of native users not really knowing what it is they need to know is dealt with.

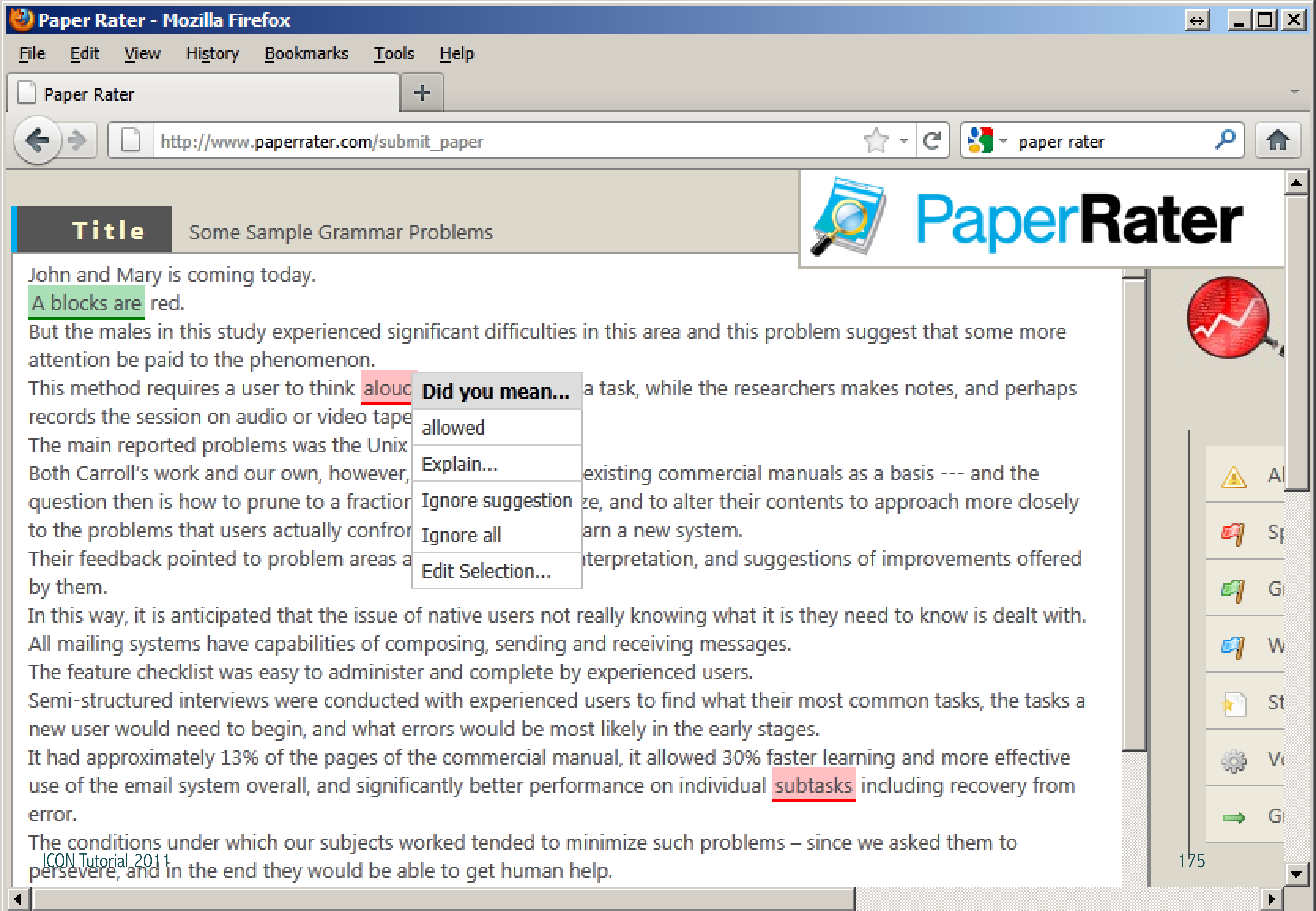
All mailing systems have capabilities of composing, sending and receiving messages.

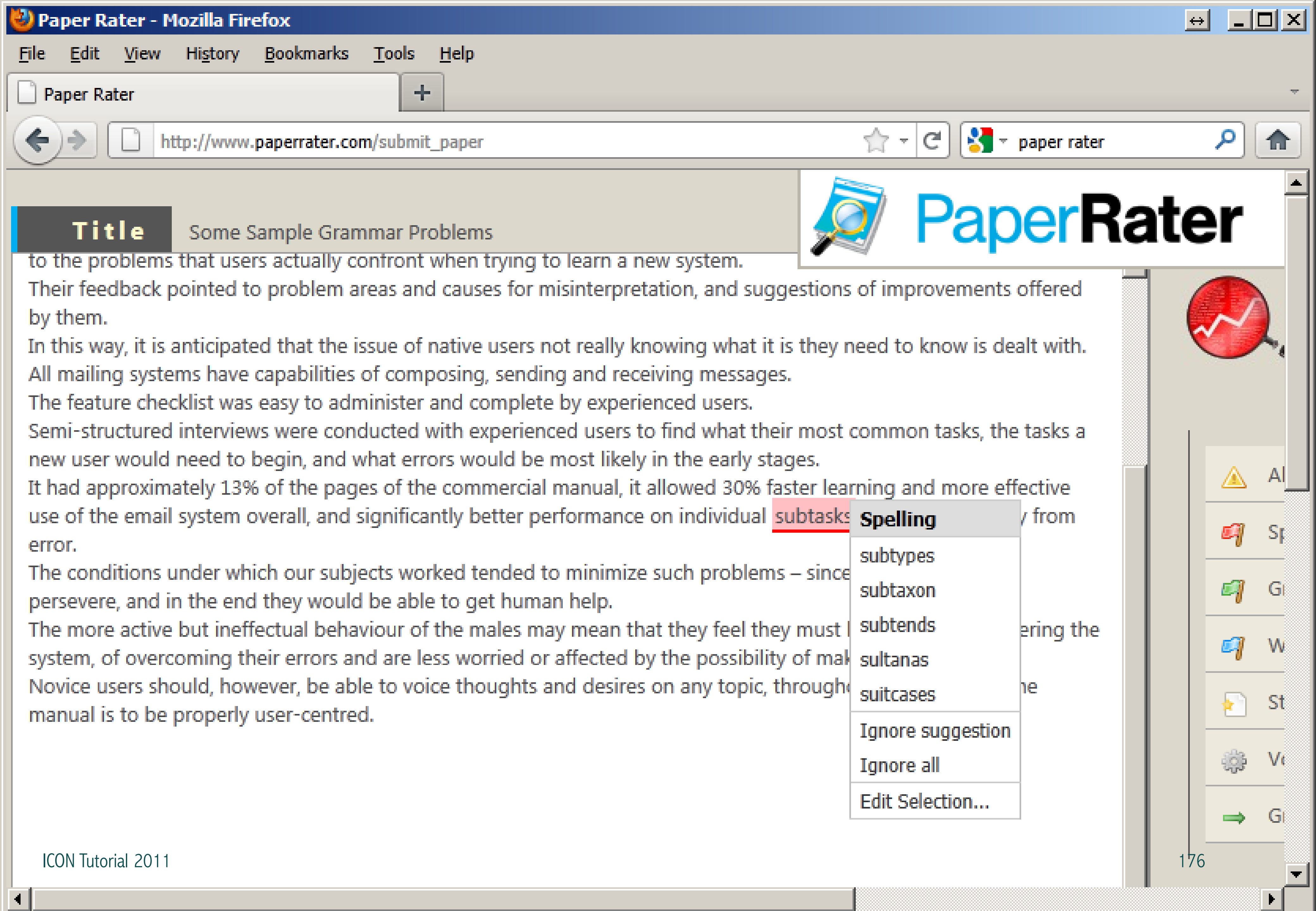
The feature checklist was easy to administer and complete by experienced users.

Semi-structured interviews were conducted with experienced users to find what their most common tasks, the tasks a new user would need to begin, and what errors would be most likely in the early stages.

It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and significantly better performance on individual subtasks including recovery from error.

The conditions under which our subjects worked tended to minimize such problems – since we asked them to persevere, and in the end they would be able to get human help.


Paper Rater - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Paper Rater
http://www.paperrater.com/submit_paper
paper rater
PaperRater
Title
Some Sample Grammar Problems
John and Mary is coming today.
A blocks are red.
But the males in this study experienced significant difficulties in this area and this problem suggest that some more attention be paid to the phenomenon.
This method requires a user to think alouc
a task, while the researchers makes notes, and perhaps
records the session on audio or video tape
The main reported problems was the Unix
Both Carroll's work and our own, however,
existing commercial manuals as a basis --- and the
question then is how to prune to a fractior
ze, and to alter their contents to approach more closely
to the problems that users actually confror
arn a new system.
Their feedback pointed to problem areas a
nterpretation, and suggestions of improvements offered
by them.
Did you mean...
allowed
Explain...
Ignore suggestion
Ignore all
Edit Selection...
In this way, it is anticipated that the issue of native users not really knowing what it is they need to know is dealt with.
All mailing systems have capabilities of composing, sending and receiving messages.
The feature checklist was easy to administer and complete by experienced users.
Semi-structured interviews were conducted with experienced users to find what their most common tasks, the tasks a new user would need to begin, and what errors would be most likely in the early stages.
It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and significantly better performance on individual subtasks including recovery from error.
The conditions under which our subjects worked tended to minimize such problems – since we asked them to persevere, and in the end they would be able to get human help.


Paper Rater - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Paper Rater
http://www.paperrater.com/submit_paper
paper rater
PaperRater
Title
Some Sample Grammar Problems
to the problems that users actually confront when trying to learn a new system.
Their feedback pointed to problem areas and causes for misinterpretation, and suggestions of improvements offered by them.
In this way, it is anticipated that the issue of native users not really knowing what it is they need to know is dealt with.
All mailing systems have capabilities of composing, sending and receiving messages.
The feature checklist was easy to administer and complete by experienced users.
Semi-structured interviews were conducted with experienced users to find what their most common tasks, the tasks a new user would need to begin, and what errors would be most likely in the early stages.
It had approximately 13% of the pages of the commercial manual, it allowed 30% faster learning and more effective use of the email system overall, and significantly better performance on individual subtasks
from
error.
The conditions under which our subjects worked tended to minimize such problems – since
persevere, and in the end they would be able to get human help.
The more active but ineffectual behaviour of the males may mean that they feel they must
ering the
system, of overcoming their errors and are less worried or affected by the possibility of mal
Novice users should, however, be able to voice thoughts and desires on any topic, through
he
manual is to be properly user-centred.
Spelling
subtypes
subtaxon
subtends
sultanas
suitcases
Ignore suggestion
Ignore all
Edit Selection...

Conclusions

- Grammar checking is hard even for humans
- Automated grammar checking is a <u>very</u> unsolved problem
- Grammar checking is not necessarily distinct from spelling checking and style checking
- Many of the problems in real texts are more complex than straightforward textbook grammar errors, and often co-occur with other errors
- There's lots to be done!

Overview

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

Outline

- Background
- Article Errors
- Preposition Errors
- Other ESL Problems
- Conclusions

Terminology

- ESL = English as a Second Language
  - Refers to non-native speakers living and speaking in a predominantly English-speaking environment
- EFL = English as a Foreign Language
  - Refers to non-native speakers studying and learning English in a non-English speaking country
- We'll generally use the term ESL to refer to both
- Apologies that this is mostly about <u>E</u>SL – there's less work in other languages …

The Problem

- Lots of people want to speak English: it is the most commonly studied second language
- Over 1 billion people speak English as a second or a foreign language
- Existing grammar checking tools are not, so far, tailored to the needs of ESL learners

ESL Errors Are Different: Bolt [1992]

- Bolt tested seven grammar-checking programs of the time against 35 sentences containing ESL errors
- Looked at from the perspective of a learner of English at a fairly low level of competence
- Conclusions:
  - ‘all of these programs fail in terms of the criteria that have been used.’
  - Expectations are encouraged that cannot be fulfilled
  - Silence on the part of a program suggests everything is ok

ESL Errors Are Different: Donahue [2001] vs Connors + Lundsford [1988]

Error	US	ESL
No comma after introductory element	1	negligible
Vague pronoun reference	2	negligible
No comma in compound sentence	3	12
Wrong word	4	2
No comma in nonrestrictive element	5	negligible
Wrong or missing inflected ends	6	6
Wrong or missing preposition	7	5
Comma splice	8	1
Possessive apostrophe error	9	negligible
Tense shift	10	negligible
Unnecessary shift in person	11	15
Sentence fragment	12	7
Wrong tense or verb form	13	4
Subject-verb agreement	14	11
Lack of comma in a series	15	negligible
Pronoun agreement error	16	negligible
Unnecessary commas with restrictive relative pronouns	17	negligible
Run on, fused sentences	18	8
Dangling, misplaced modifier	19	negligible
Its, it's confusion	20	negligible

ESL Errors Are Different

- Frequent error types for ESL speakers are negligible in the native speaker population:

Error	US	ESL
Missing words	negligible	3
Capitalization	negligible	9
Wrong pronoun	negligible	16
a, an confusion	negligible	14
Missing article	negligible	17
Wrong verb form	negligible	10
No comma before etc.	negligible	13

- Half of the ten most frequent error types made by native speakers are negligible in the writing of the ESL population

Errors in the Cambridge Learners Corpus

Rank	Error Type	Prop	Example sentence
1	Content word choice error	0.199	We need to deliver the merchandise on a daily **base/basis*.
2	Preposition error	0.134	Our society is developing **in/at* high speed.
3	Determiner error	0.117	We must try our best to avoid **the/a* shortage of fresh water.
4	Comma error	0.093	However, */, I'll meet you later.
5	Inflectional morphology	0.074	The women **weared/wore* long dresses.
6	Wrong verb tense	0.067	I look forward to **see/seeing* you.
7	Derivational morphology	0.049	It has already been **arrangement/arranged*.
8	Pronoun	0.042	I want to make **me/myself* fit.
9	Agreement error	0.040	I **were/was* in my house.
10	Run-on Sentence	0.040	The deliver documents to them *they* provide fast service.
11	Idiomatic Collocation and word order	0.039	The latest issue **the magazine of/of the magazine* …
12	Confused words	0.019	I want to see the **personal/personnel* manager.
13	Conjunction error	0.017	I want to see you **and/so* that you can help me.
14	Words split with a space or joined	0.014	I organize sports **everyday/every day*. It is also my **life style/lifestyle*.
15	Apostrophe error (not including *it/it's* confusions)	0.013	We are all **sport's/sports* lovers.
16	Hyphenation error	0.013	It is a nourishing **low cost/low-cost* meal.
17	Sentence fragment or two sentences that are joined	0.008	I'm going to get another one *. *Because/because* the old one broke.
18	Quantifier error	0.007	It doesn't give them too **much/many* problems.
19	Other punctuation error	0.004	When are you leaving *./?
20	Negation formation	0.001	We **have not/do not have* any time.

Common ESL Errors

- The most difficult aspects of English for ESL learners are:
  - Definite and indefinite articles
  - Prepositions
- Together these account for 20–50% of grammar and usage errors
- [But: spelling errors are much more common, and incorrect word choice is as problematic as article and preposition errors.]

What Causes the Problem?

- Articles :
  - Not present in all L1s
  - Correct article choice is very subtle and depends on a complex discourse and real world knowledge factors
- Prepositions:
  - Behaviour appears very idiosyncratic and unpredictable

Article Errors in the CLC by L1

L1	Has Articles	Proportion
Russian	No	0.186
Korean	No	0.176
Japanese	No	0.159
Chinese	No	0.125
Greek	Yes	0.087
French	Yes	0.081
Spanish	Yes	0.070
German	Yes	0.053

Proportion of sentences with one or more article errors

Preposition Errors in the CLC by L1

L1	Proportion
Greek	0.149
Spanish	0.139
Korean	0.128
Chinese	0.122
French	0.121
Japanese	0.118
German	0.100
Russian	0.095

Proportion of sentences with one or more preposition errors

A Note on Data

- The field has been hamstrung by the privately held nature of many learner corpora
- Two welcome changes:
  - The NUS Corpus of Learner English
  - The Cambridge Learner Corpus FCE Dataset
- Also the much smaller HOO dataset

NUCLE: The NUS Corpus of Learner English

- 1400 essays written by University students at the National University of Singapore
- Over 1M words annotated with error tags and corrections
- See http://nlp.comp.nus.edu.sg/corpora

NUCLE: The NUS Corpus of Learner English

Standoff annotation:

```
<MISTAKE start_par="4" start_off="194" end_par="4" end_off="195">
   <TYPE>ArtOrDet</TYPE>
   <CORRECTION>an</CORRECTION>
</MISTAKE>
```

The CLC FCE Dataset

- A set of 1,244 exam scripts written by candidates sitting the Cambridge ESOL First Certificate in English (FCE) examination in 2000 and 2001
- Annotated with errors and corrections
- A subset of the much larger 30M-word Cambridge Learner Corpus
- See http://ilexir.co.uk/applications/clc-fce-dataset/

The CLC FCE Dataset

Inline annotation:

- Because <NS type="UQ"><i>all</i></NS> students in <NS type="MD"><c>the</c></NS> English class are from all over the world …

The HOO Dataset

- HOO – Helping Our Own – aims to marshall NLP technology to help non-native speakers write ACL papers
- Very small corpus (~36K words) annotated with errors and corrections
- Evaluation software also freely available
- See http://www.clt.mq.edu.au/research/projects/hoo/

The HOO Dataset

Stand-off and inline annotation both available:

- In our experiments, pseudo-words are fed into <edit type="MD"><empty/><corrections><correction>the</correction></corrections></edit> PB-SMT pipeline.
- <edit index="1005-0016" type="MD" start="871" end="871" >
  <original><empty/></original>
  <corrections>
  <correction>the </correction>
  </corrections>
  </edit>

Outline

- Background
- Article Errors
- Preposition Errors
- Other ESL Problems
- Conclusions

Article Errors

- The Problem
- Early Rule-based Approaches
- Knight and Chandler [1994]
- Han et al [2006]
- De Felice and Pulman [2008]

Why is Article Choice Hard?

- Basic problem for speakers of languages that do not use articles:
  - choose between a/an, the, and the null determiner
- The bottom line: it comes down to context
  - I was eating a cake.
  - I was eating the cake.
  - I was eating cake.

Features Impacting Article Choice: Countability

- Count nouns take determiners:
  - I read the paper yesterday.
- Mass nouns don't take determiners:
  - We generally write on paper.
- But the universal grinder and the universal packager [Pelletier 1975] are always available:
  - There was dog all over the road.
  - Could we have just one rice please?

Features Impacting Article Choice: Countability

- Semi-idiomatic forms:
  - I looked him in the eye.
  - *I looked him in an eye.

Features Impacting Article Choice: Syntactic Context

- ✓ I have knowledge.
- ✗ I have a knowledge.
- ✓ I have knowledge of this.
- ✗ I have a knowledge of this.
- ✓ I have a knowledge of English.

Features Impacting Article Choice: Discourse Factors

- Stereotypically, entities are introduced into a discourse using an indefinite determiner and subsequently referred to using a definite determiner
  - I saw <u>a man</u> at the bus stop. … <u>The man</u> was crying.
- But not always:
  - A bus turned the corner. <u>The driver</u> was crying.
  - I went to <u>the beach</u> yesterday.

Features Impacting Article Choice: World Knowledge

- He bought a Honda.
- He bought Honda.

Article Errors

- The Problem
- Early Rule-based Approaches
- Knight and Chandler [1994]
- Han et al [2006]
- De Felice and Pulman [2008]

Early Work: Article Insertion in Machine Translation

- The Problem:
  - Machine translation of languages like Japanese or Russian into English is difficult because the source language doesn't contain articles

Murata and Nagao [1993]: Hand-Crafted Rules

- When a noun is modified by a referential pronoun (KONO(this), SONO(its), ...) then {indefinite(0, 0), definite(1, 2), generic(0, 0)}
- When a noun is accompanied by a particle (WA), and the predicate has past tense, then {indefinite(1, 0), definite(1, 3), generic(1, 1)}
- When a noun is accompanied by a particle (WA), and the predicate has present tense, then {indefinite(1, 0), definite(1, 2), generic(1, 3)}
- When a noun is accompanied by a particle HE(to), MADE(up to) or KARA(from), then {indefinite(1, 0), definite(1, 2), generic(1, 0)}
- ... 84 heuristics in total

Article Errors

- The Problem
- Early Rule-based Approaches
- Knight and Chandler [1994]
- Han et al [2006]
- De Felice and Pulman [2008]

Knight and Chandler [1994]: A Data-Driven Method for Post-Editing

- **General aim:**
  - To build a post-editing tool that can fix errors made in a machine translation system
- **Specific task:**
  - Article insertion: *a*, *an* or *the*

Knight and Chandler [1994]: Before and After

Stelco Inc. said it plans to shut down three Toronto-area plants, moving their fastener operations to leased facility in Brantford, Ontario.

Company said fastener business "has been under severe cost pressure for some time." Fasteners, nuts and bolts are sold to North American auto market.

Company spokesman declined to estimate impact of closures on earnings. He said new facility will employ 500 of existing 600 employees. Steelmaker employs about 16,000 people.

Stelco Inc. said it plans to shut down three Toronto-area plants, moving their fastener operations to **a** leased facility in Brantford, Ontario.

The company said **the** fastener business "has been under severe cost pressure for some time." **The** fasteners, nuts and bolts are sold to **the** North American auto market.

A company spokesman declined to estimate **the** impact of **the** closures on earnings. He said **the** new facility will employ 500 of **the** existing 600 employees. **The** steelmaker employs about 16,000 people.

Knight and Chandler [1994]: The General Idea

The steps:

- Take newspaper-quality English text
- Remove articles
- Re-insert automatically
- Compare results with the original text

Assumptions:

- NPs are marked as singular or plural
- Locations of articles already marked so it's a binary choice between *the* and *a/an*.

Knight and Chandler [1994]: Baseline

- In 40Mb of Wall Street Journal text:

  a = 28.2%

  an = 4.6%

  the = 67.2%

- So 67% is a good lower-bound
- Upper-bound:
  - Human subjects performed with accuracy of 94%-96%

Knight and Chandler [1994]: Baselines

	Human	Machine
Random	50%	50%
Always guess *the*	67%	67%
Given head noun + premodifiers (the 'core NP')	79-80%	
Given core NP + 2 words either side	83-88%	?
Given full context	94-96%	

Knight and Chandler [1994]: Approach

- Characterize NPs via sets of features then build a decision tree to classify
- Lexical features:
  - ‘word before the article is *triple*’
- Abstract features:
  - ‘word after the head noun is a past tense verb’
- 400k training examples and 30k features; features with less than 4 instances discarded

Knight and Chandler [1994]: Performance

- On 1600 trees for the 1600 most frequent head nouns (covering 77% of test instances):
  - 81% accuracy
- Guess *the* for the remaining 23% of test instances
  - 78% accuracy overall

Article Errors

- The Problem
- Early Rule-based Approaches
- Knight and Chandler [1994]
- Han et al [2006]
- De Felice and Pulman [2008]

Han et al [2006]:
A MaxEnt Approach to Article Selection

- Basic Approach:
  - A maximum entropy classifier for selecting amongst *a/an*, *the* or the null determiner
  - Uses local context features such as words and PoS tags

Han et al [2006]: Contrasts with Earlier Work

- More varied training corpus: a range of genres instead of just one source:
  - 721 text files, 31.5M words
  - 10th thru 12th grade reading level
- Much larger training corpus: 6 million NPs (15x larger than Knight and Chandler's)
  - Automatically PoS tagged + NP-chunked
- The use of a maximum entropy classifier

Han et al [2006]: Training Results

- 6M NPs in training set
- 390k features
- Baseline = 71.84% (frequency of null determiner)
- Four-fold cross validation
  - performance range 87.59% to 88.29%
  - Average 87.99%

Han et al [2006]: Effectiveness of Individual Features

Feature	% Correct
Word/PoS of all words in NP	80.41
Word/PoS of w(NP-1) + Head/PoS	77.98
Head/PoS	77.30
PoS of all words in NP	73.96
Word/PoS of w(NP+1)	72.97
Word/PoS of w(NP[1])	72.53
PoS of w(NP[1])	72.52
Word/PoS of w(NP-1)	72.30
PoS of Head	71.98
Head's Countability	71.85
Word/PoS of w(NP-2)	71.85
Default to null determiner	71.84

Han et al [2006]: Effectiveness of Individual Features

- Best feature: Word/PoS of all words in NP
  - Ok if you have a large enough corpus!
- Second best: W(NP-1) + Head
  - Eg 'in summary'

Han et al [2006]: Impact of Training Set Size

Number of NPs	% Correct
75000	83.03
150000	83.49
300000	84.92
600000	85.75
1200000	86.59
2400000	87.27
4800000	87.92
6000000	97.99

Han et al [2006]: Impact of Frequency of Head Noun

Frequency of Head Noun	% Correct
1	73.6
5	73.6
10	76.0
50	78.5
100	79.6
500	80.7
1000	81.9
5000	82.4
10000+	86.3

Han et al [2006]: Accuracy by Head Noun Type

Syntactic Type of Head	% Correct
Singular Noun	80.99
Plural Noun	85.02
Pronoun	99.66
Proper Noun, Singular	90.42
Proper Noun, Plural	82.05
Number	92.71
Demonstrative Pronoun	99.70
Other	97.81

Han et al [2006]: Applying the Model to TOEFL Essays

- Model retrained only on NPs with a common head noun
  - Baseline = frequency of null determiner = 54.40%
  - Training set kept at 6M instances by adding more data
  - Average accuracy = 83.00%
- Model applied to 668 TOEFL essays w 29759 NPs
  - Subset of NPs classified by two annotators
  - Agreement on 98% of cases with kappa = 0.86
  - One article error every 8 NPs

Han et al [2006]: Some Examples

Above all, I think it is good for students to share <u>room</u> with others.

- Human: missing *a* or *an*
- Classifier: 0.841 *a/an*; 0.143 *the*; 0.014 zero

Those excellent hitters began practicing <u>the baseball</u> when they were children, and dedicated a lot of time to become highly qualified.

- Human: superfluous determiner
- Classifier: 0.103 *a/an*; 0.016 *the*; 0.879 zero

Han et al [2006]: Results on TOEFL Essays

- 79% of errors in test set correctly detected
- Many false positives, so precision only 44%
- Decisions often borderline:
  - The books are assigned by professors.
  - Marked by annotators as correct, model predicts *the* (0.51) and null (0.49)

Han et al [2006]: Sources of Error

- Model performs poorly on decision between *a* and *the*
  - Probably due to the need for <u>discourse information</u>
- So, new feature: has the head noun appeared before, and if so, with what article?
  - No significant effect on performance
- Error analysis suggests this is due to more complex discourse behaviour:
  - A student will not learn if she hates <u>the teacher</u>.
  - … the possibilities that <u>a scholarship</u> would afford …

Article Errors

- The Problem
- Early Rule-based Approaches
- Knight and Chandler [1994]
- Han et al [2006]
- De Felice and Pulman [2008]

De Felice and Pulman [2008]: Richer Syntactic and Semantic Features

- Basic Approach:
  - As in Han et al [2006], a maximum entropy classifier for selecting amongst *a/an*, *the* or the null determiner
  - Use a richer set of syntactic and semantic features

De Felice and Pulman [2008]: Main Features

Feature	Value
Head Noun	‘apple’
Number	Singular
Noun Type	Count
Named Entity?	No
WordNet Category	Food, Plant
Prepositional Modification?	Yes, ‘on’
Object of Preposition?	No
Adjectival Modification?	Yes, ‘juicy’
Adjectival Grade	Superlative
POS±3	VV, DT, JJS, IN, DT, NN

Example: Pick the juiciest apple on the tree.

De Felice and Pulman [2008]: Additional Features

- Whether the noun is modified by a predeterminer, possessive, numeral and/or a relative clause
- Whether it is part of a ‘there is ...’ phrase

De Felice and Pulman [2008]: Performance

- Trained on British National Corpus
  - 4,043,925 instances
- Test set of 305,264 BNC instances
- Baseline = 59.83% (choose null)
- Accuracy = 92.15%

De Felice and Pulman [2008]: Comparative Performance on L1 Data

Author	Accuracy
Baseline	59.83%
Han et al 2006	83.00%
Gamon et al 2008	86.07%
Turner and Charniak 2007	86.74%
De Felice and Pulman 2008	**92.15%**

De Felice and Pulman [2008]: Results on Individual Determiners

	% of Training Data	Precision	Recall
a	9.61% (388,476)	70.52%	53.50%
the	29.19% (1,180,435)	85.17%	91.51%
null	61.20% (2,475,014)	98.63%	98.79%

- **The indefinite determiner is less frequent and harder to learn**

De Felice and Pulman [2008]: Testing on L2 Text

- 3200 instances extracted from the CLC
  - 2000 correct
  - 1200 incorrect
- Accuracy on correct instances: 92.2%
- Accuracy on incorrect instances: < 10%
- Most frequent incorrect usage is a missing determiner
  - Model behaviour influenced by skew in training data
- Also problems in extracting NLP features from L2 data

Outline

- Background
- Article Errors
- Preposition Errors
- Other ESL Problems
- Conclusions

The Prevalence of Preposition Errors

L1	Proportion
Greek	0.149
Spanish	0.139
Korean	0.128
Chinese	0.122
French	0.121
Japanese	0.118
German	0.100
Russian	0.095

Proportion of sentences in the CLC with one or more preposition errors

Prepositions Have Many Roles in English

- They appear in adjuncts:
  - In total, I spent $64 million dollars.
- They mark the arguments of verbs:
  - I'll give ten cents to the next guy.
- They figure in phrasal verbs:
  - I ran away when I was ten.
- They play a part in idioms:
  - She talked down to him.

Negative Transfer

- Many prepositions have a most typical or frequent translation
  - Eg: *of* in English to *de* in French
- But for many prepositions there are multiple translational possibilities
  - ESL speakers can easily choose the wrong one
  - Eg: driving *in* a high speed

Prepositions in English

- English has over 100 prepositions, including some multiword prepositions and a small number of postpositions
- The 10 most frequent account for 82% of the errors in the CLC

Preposition Selection in Well-Formed Text

Citation	Approach	Training Corpus	Testing Corpus	Performance
Lee and Seneff (2006)	Parse Ranking, using Collins (1999)'s parser	10,369 Transcripts of flight reservation data	317 sentences from transcripts of flight reservation data	P=88%, R=78%
Chodorow et al. (2007)	Maximum Entropy Classifier and Heuristic Rules, token context, part-of-speech context, chunk information	SJM and MetaMetrics: 7M cases	MetaMetrics: 18.2K cases	69% accuracy
De Felice and Pulman (2007)	Voted Perceptron, part-of-speech context, parse information, semantic information	BNC: 10-fold xval	BNC subset, 10k sentences	76% accuracy
De Felice and Pulman (2008)	Maximum Entropy Classifier, part-of-speech context, parse information, semantic information	BNC: 9M cases	BNC: 536.2K cases	70% accuracy
Tetreault and Chodorow (2008b)	Maximum Entropy Classifier and Heuristic Rules, token context, part-of-speech context	SJM and MetaMetrics: 10M cases (plus NANTC and Encarta/Reuters)	WSJ, Encarta/Reuters (1.4M cases)	90% accuracy (WSJ), 79% accuracy (Encarta/Reuters)
Gamon et al. (2008)	Decision Tree and Language Model, token context, part-of-speech context	Encarta, Reuters, etc.	Encarta/Reuters 1.4M cases	combined accuracy =77% (presence/absence = 91%, choice=62%)
Bergsma et al. (2009)	Google N-gram corpus approach (log of counts)	NYT: 1M cases	NYT: 10K cases	75% accuracy

Preposition Error Detection on Learner Data

Citation	Approach	Training Corpus	Testing Corpus	Performance
Eeg-Olofsson and Knutt (2003)	Heuristic Rules, parse information	n/a	40 cases from Swedish learner essays	11/40 correct
Tetreault and Chodorow (2008a)	Maximum Entropy Classifier and Heuristic Rules, token context, part-of-speech context	SJM and MetaMetrics: 7M cases	TOEFL: 8.2K cases	P=84%, R=19%
De Felice and Pulman (2009)	Maximum Entropy Classifier, token context, part-of-speech context, semantic information	BNC: 9M cases	CLC: 1116 incorrect cases, 5753 correct cases	P=42% and R=35% on incorrect cases, accuracy 69% on correct cases
Hermet et al. (2008)	Web-counts method	WWW	133 French Learner sentences	70% accuracy on error correction task
Tetreault and Chodorow (2009)	Web-counts method (Region Counts Approach)	WWW	TOEFL: 518 cases for 5 constructions	n/a
Gamon (2010)	Maximum Entropy and LM, token context, part-of-speech context	2.5M sentences of well-formed text; LM (Gigaword); CLC for meta-classifier	CLC (208.7K sentences/19.7K errors)	Auto: P=35%, R=22%; Manual verification: 6K sentences, P=85%
Han et al. (2010)	Maximum Entropy, token context, part-of-speech context, parse information	Chungdahm: 978,000 error-annotated cases	Chungdahm: 1,000 cases	Detection: P=93%, R=15%; Correction: P=82%, R=13%

Upcoming Shared Task

Advertisement!

- HOO 2012 at the Building Educational Applications Workshop at NAACL 2012
- Preposition and Determiner Error Correction
- See www.correcttext.org/hoo2012
- Schedule:
  - Friday 27th January: Registration opens
  - Friday 6th April: Test data for evaluation released
  - Friday 13th April: Deadline for submissions for evaluation.
  - Friday May 4th: Team reports deadline for proceedings

Outline

- Background
- Article Errors
- Preposition Errors
- Other ESL Problems
- Conclusions

Collocations

- Conventional combinations that are preferred over other equally syntactically and semantically valid combinations
  - Adj + Noun: *stiff breeze* vs *rigid breeze*
  - Verb + Noun: *hold an election* vs *make an election*
  - Noun + Noun: *movie theatre* vs *film theatre*
  - Adverb + Verb: *thoroughly amuse* vs *completely amuse*

Collocations

- Computational approaches generally make use of distributional differences for detecting and correcting errors
- Same general approach as in articles and prepositions:
  - Choose preferred form from a set of alternatives
  - But: the confusion set is potentially much larger
- Solution:
  - Constrain the space by selecting alternatives with a similar meaning
- See work on automatic thesaurus construction [eg Lin 1998]

Verb Form Errors

Error Type	Example
Subject-Verb Agreement	He have been living here since June.
Auxiliary Agreement	He has been live here since June.
Complementation	He wants live here.

- **See Lee and Seneff [2008] for a method based on detecting specific irregularities in parse trees.**

Outline

- Background
- Article Errors
- Preposition Errors
- Other ESL Problems
- Conclusions

Conclusions

- The provision of assistance to ESL learners is clearly a significant market
- Technology is at a very early stage, focussing on specific subproblems
- Measurable progress has been hampered by the unavailability of shared data sets, but this is changing

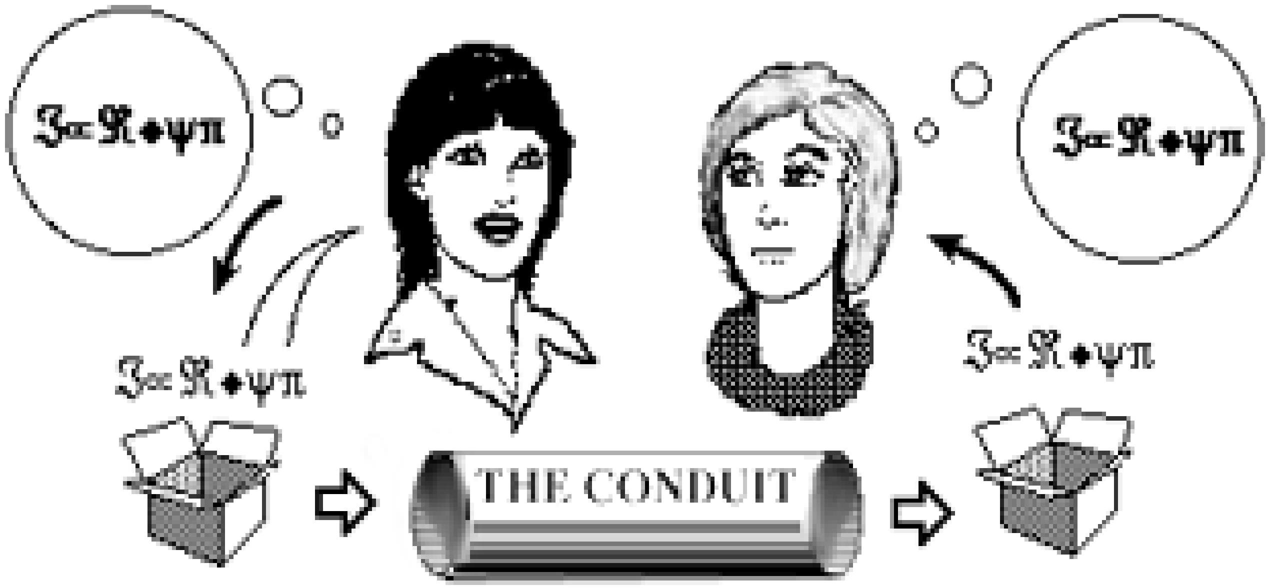
Overview

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

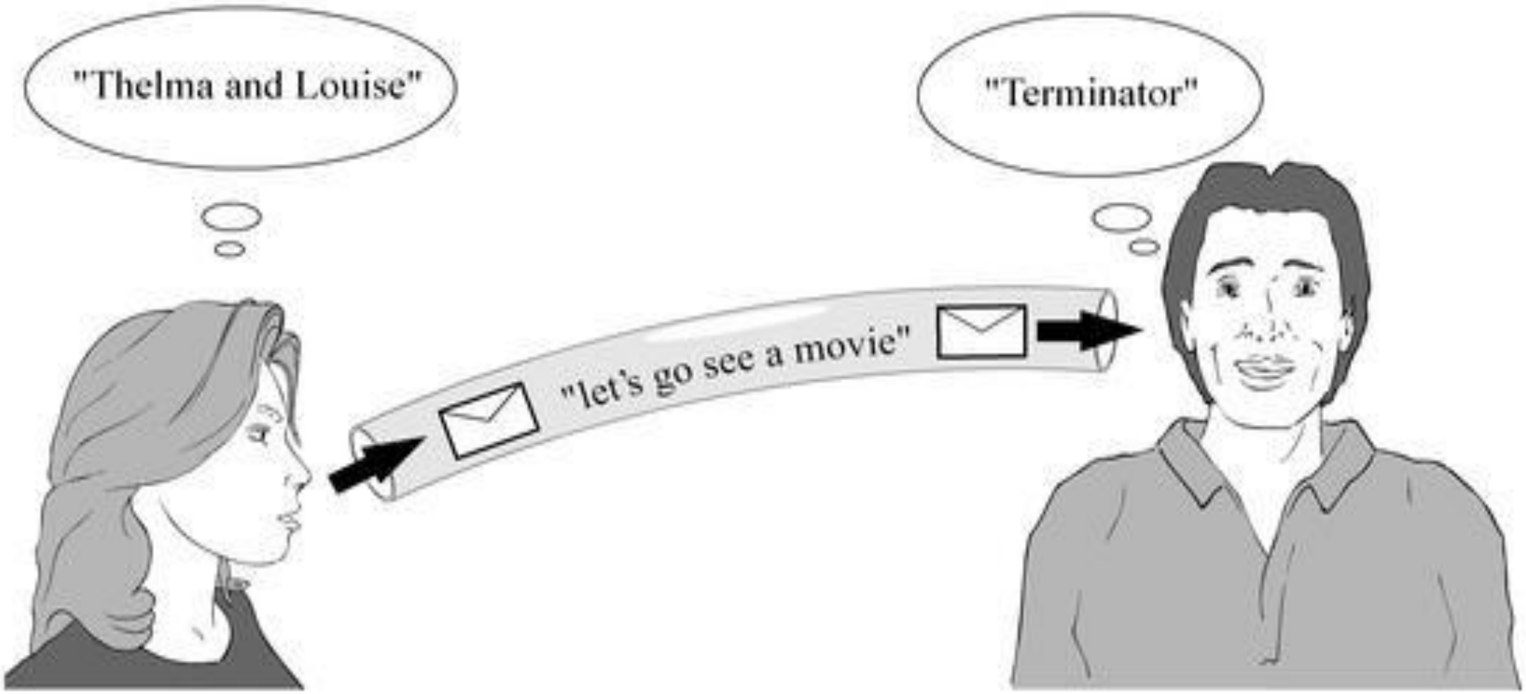
Outline

- The Nature of the Writing Process
- Help with Revision

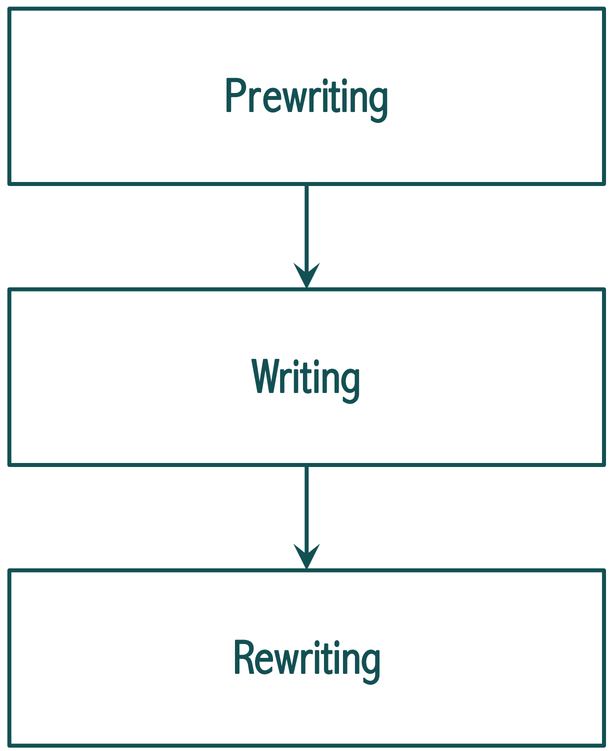
The Conduit Metaphor #1



The Conduit Metaphor #2

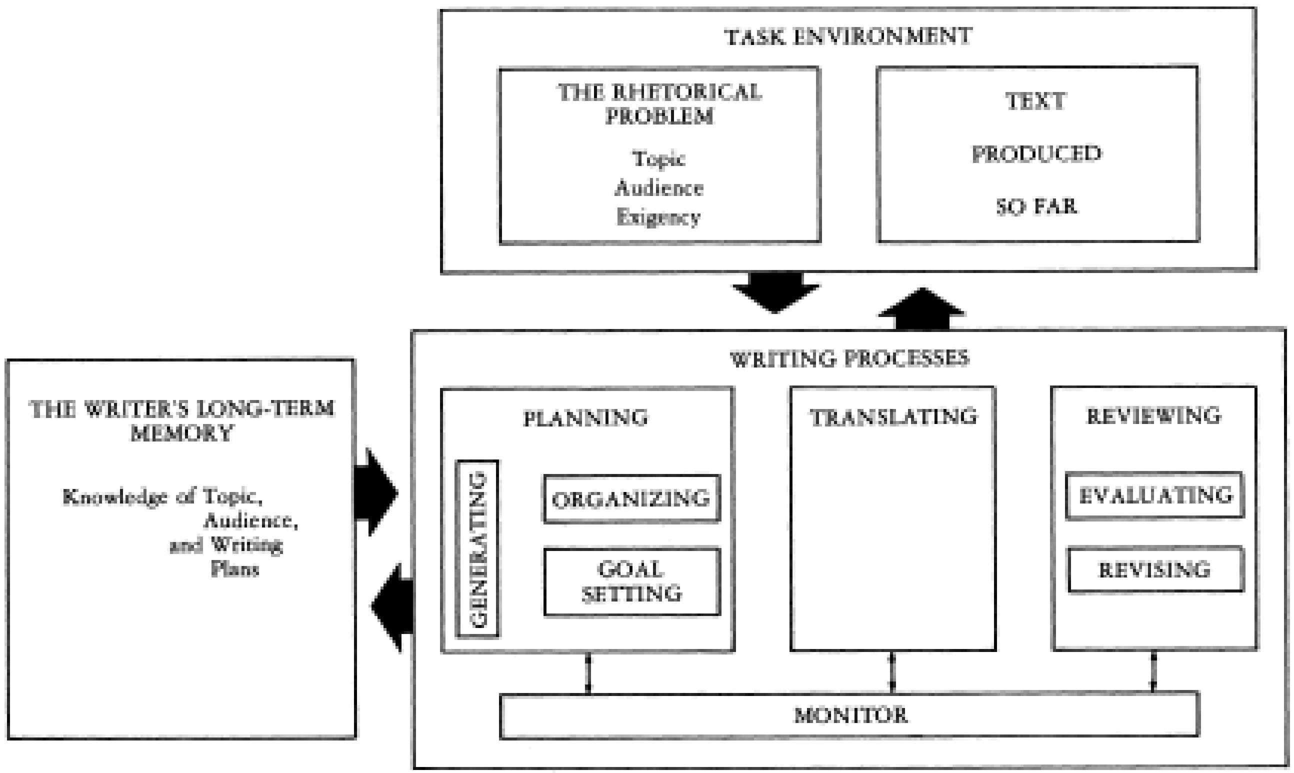


A Stage Model of the Writing Process



Rohman 1965

A Cognitive Process Model



Flower and Hayes 1981

Outline

- The Nature of the Writing Process
- Help with Revision

The Nature of Revision

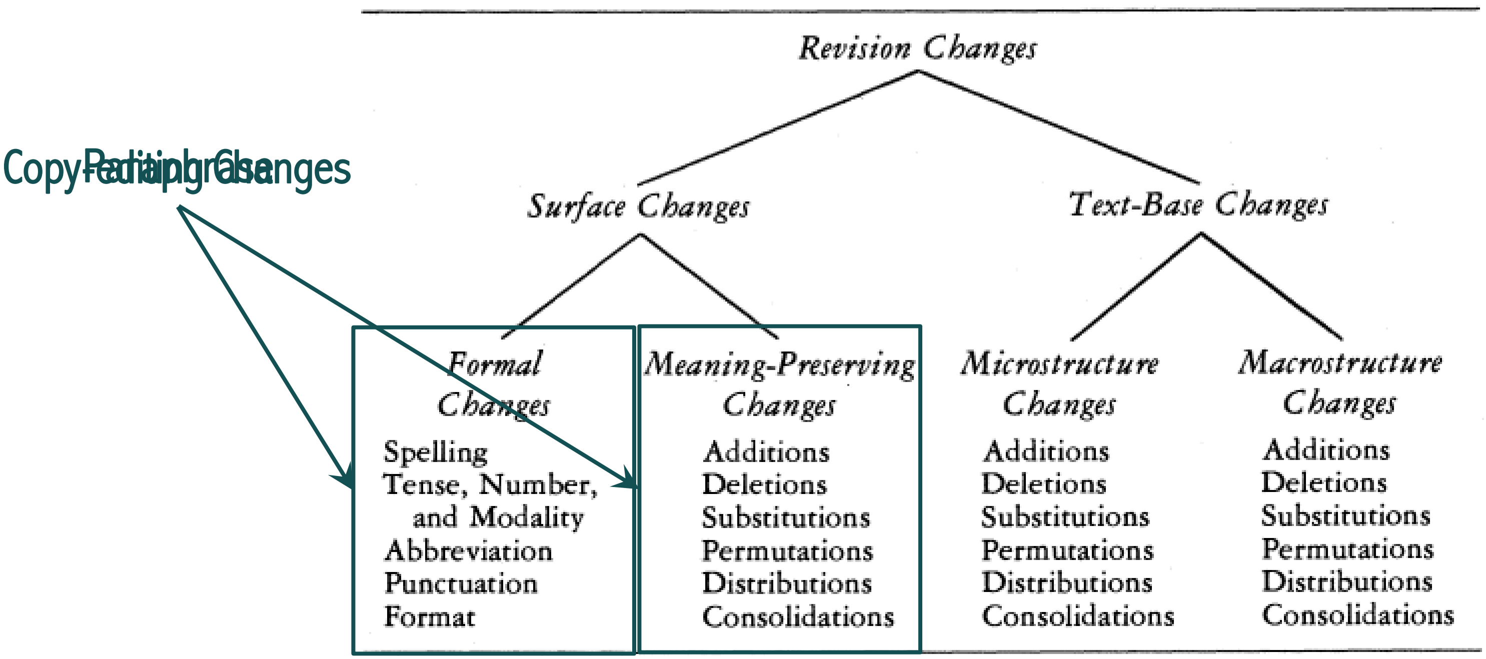


Figure 1. A Taxonomy of Revision Changes

Faigley and Witte 1981

Meaning-Preserving Changes: Additions

Additions make explicit what can be inferred:

- you pay two dollars $\rightarrow$ 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 $\rightarrow$ out-of-the-way places

Meaning-Preserving Changes: Permutations

<u>Permutations</u> rearrange material, possibly with substitutions:

- springtime means to most people
  $\rightarrow$ 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

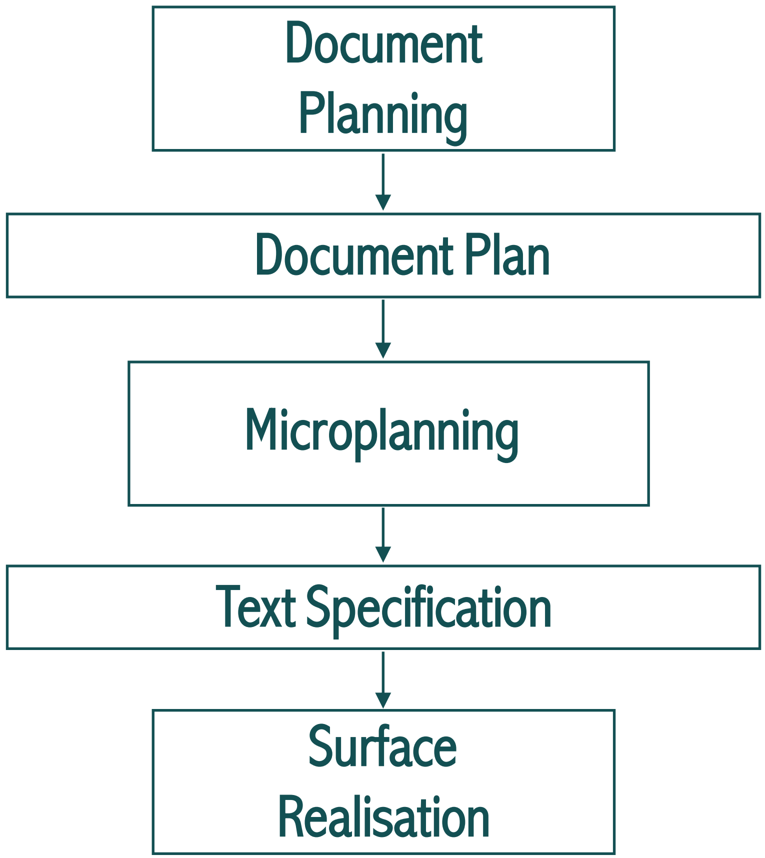
Consolidations move material from multiple units into a single unit:

- 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.

Are These Revisions Automatable?

- The relevant ideas are already a focus in various sub-areas of NLP:
  - Paraphrase, Text Simplification and Lexical Simplification
  - Recognizing Textual Entailment
  - Surface Realisation
  - Sentence Aggregation

A Pipelined Architecture for NLG



Tasks and Architecture in NLG

Tasks	Stage
• Content determination • Discourse planning	Document Planning
• Sentence aggregation • Lexicalisation • Referring expression generation	Micro Planning
• Syntax + morphology • Orthographic realization	Linguistic Realization

The Nature of Revision

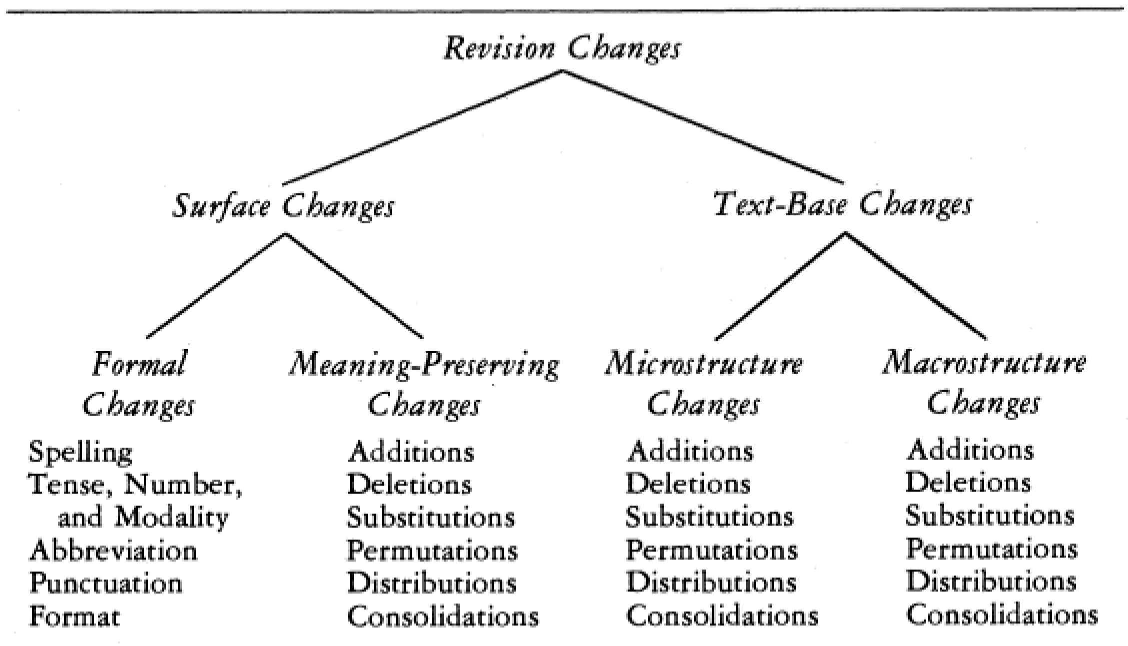


Figure 1. A Taxonomy of Revision Changes

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

Meaning Changes

- These are the focus of NLG research in
  - Content Determination
  - Text Structuring

The State of the Art and Where We Might Go

- Existing tools are concerned with surface revisions, and even then primarily with formal changes
- But: we can conceive of machine assistance being provided for every aspect of revision

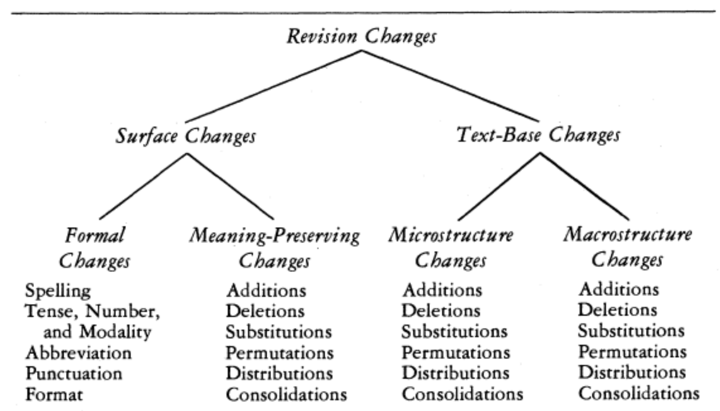


Figure 1. A Taxonomy of Revision Changes

- We can also conceive of machine assistance being provided for other stages of the writing process

Overview

- Introduction: The Need
- Spell Checking
- Grammar Checking
- Helping Non-Native Speakers
- Beyond Spelling and Grammar Checking
- Conclusions

Conclusions

- Current technology only scratches the surface in terms of the kinds of support we would like to give to authors
- Many aspects of NLP technology can be pressed into service to support authors
- NLG techniques provide a rich source of ideas for how to build symbiotic systems that take advantage of the knowledge and capabilities of both human and machine

Who Today's Main Players Are

- Google
- Microsoft
- Educational Testing Service
- Activities around the University of Cambridge

Finding Out More

- ACL Workshops on Innovative Use of NLP for Building Educational Applications: 2012 will be the seventh in the series
- Relevant material often found in journals outside the normal 'ACL space':

  CALICO Journal
  College Composition and Communication
  Computers and Composition
  Computer Assisted Language Learning
  Journal of Second Language Writing

Writing Assistance in the Future?