

CSE6390 3.0 Special Topics in AI & Interactive Systems II
Introduction to Computational Linguistics
Instructor: Nick Cercone – 3050 CSEB – nick@cse.yorku.ca
Tuesdays,Thursdays 10:00-11:30 – South Ross 104
Fall Semester, 2010

Paper Writing

A Story

Generations ago a minister in a Scottish Highland parish was mystified by the behaviour in church of the three sons of the local doctor. Though the boys were at the difficult ages of 5, 7, and 10 years, they sat entranced through the longest sermon. Such concentration is not common among boys of that age, and the minister pressed the doctor to reveal how this miracle was repeatedly performed. The doctor was reluctant to disclose how it was done but finally gave way and told the minister what was happening. The doctor took to church each Sunday a bar of chocolate. When the sermon began he gave it to the eldest boy and whenever the minister came out with one of his favourite phrases, *that is true* or *everyone will agree* or *all right thinking people*, the boy would hand the bar to his brother. Up and down it would go until the benediction was pronounced, when the boy who was holding it was allowed to eat it. This simple story illustrates a fact, which we all know, but do not much like admitting, that *we have habits of speech, and of writing too, of which we are so little aware that we sometimes seem to be their prisoner.*

A careful student of literary habits could not only distinguish an American from an English or Canadian education, but he could go much further – to that of author attribution.

An Example

In 1939 Andrew Q. Morton published a book entitled *Who Was Socrates?* Which was followed in 1940 by *The Genesis of Plato's Thought*. Morton argued that both Socrates and Plato were much more closely involved in the politics of their time than scholars had imagined, and that the work of both men was, in substance, an apologia for Greek conservatism. In working out the evidence, Morton relied on the *Seventh Letter* which he took to be Plato's political apology, just as the *Apology* of Plato was that of Socrates.

At that time, these books seemed heretical to Classical scholars.

In relying heavily on the Seventh Letter of Plato, Morton was in line with the best scholarly opinion of the time. Later the Seventh Letter came under heavy attack and its authenticity was sharply challenged! If the Letter proved to be spurious or forged, then the thesis of Morton's book would have been undermined.

Morton decided to put the Seventh Letter on the computer to examine the skeletal structure of the language. A statistical analysis revealed that if the Apology of Socrates was Platonic, the Seventh Letter must be rejected and that, in fact, the Seventh Letter was not homogeneous. – Very bad news indeed!

Other Classics/Language Studies

- Study of article use
- The Distribution of syllables per word in texts
- Homogeneity within works of authors
- Syllable counts for nouns
- Lexicostatistical models for testing affinities between languages of unknown origin
 - Taking phonologies of languages into account
 - Taking diagram regularities into account
 - Taking phoneme relative frequencies into account
- Sir Arthur Eddington's lectures – [the monkeys](#)

Seven Deadly Sins in Journal Publishing

- Chapter 11 in *The Great Eskimo Vocabulary Hoax and other Irreverent Essays on the Study of Language* by Geoffrey Pullum, University of Chicago Press, 1991.
- Methodology – Pullum’s research team surveyed no less than 7 journals – a large and wide ranging survey!
- Why is this a large and wide ranging survey, you ask?
 - The median number of speakers on whom the entire corpus of examples in an English syntax paper is checked before publication, including its author, is zero.
 - The median number of children used in an acquisition study is one
 - The total number of experimental subjects used in Lieberman’s study of the perception of intonation by linguists (Word 21, 40-54) was two
 - The total number of Russian forms in Chomsky’s classic presentation of the Halleian argument against phonemics (Current issues in linguistic theory, section 4.3) is four.
- Thus **seven** is a big number to linguists!.

No No’s

- Omitting the date of receipt and/or acceptance of a published article
- Failing to announce in an issue its month of publication
- Concealing the mailing address of an author
- Lacking a contents page for each issue on its cover
- Neglecting to print on the spine of each issue an indication of the page numbers covered in it
- Forgetting to include a statement of reference details for each article on its first page
- Hiding footnotes at the end of an article

Writing Tidbits

- Avoid False Starts (we just had two)
- Have a novel idea that has some significance
- Simplicity and Elegance go hand in hand
- Sometimes less is more
- Avoid Pronouns
- Use short direct English sentences
- Pictures and Examples (sometimes notation) help but ...
 - Reference them in text
 - Explain them
 - Use pictures and examples (sometimes notation) to clarify
- Reference by (sub)section number or equation number
- Always follow “this” with a noun
- Avoid lengthy definite descriptions, reference to a numbered passage is better
- After paper completion and independent reader “authentication” have an independent reader “authenticate” the paper (again...)
- Check spelling then check spelling (watch “you” vs. “your” and so on – that spell checkers don’t check)
- Be direct, active voice, to the point.
- Show your excitement, if you don’t, why should the reader care?
- Strive for quality – people understand things differently

IMRAD – purported to be the only way to write a scientific article

- Introduction
- Methodology
- Results
- Analysis
- Discussion

But – a Simple Approach to Thesis Writing

- Introduction - Motivation, Goals / Objectives, Contributions, Outline of the paper (optional)
- Related Work – critical review, identify opportunities for more research
- Theory / Solution / Program / Problem

- Implementation / Formalism - not every paper has or needs an implementation
- Results and Evaluation
- Conclusions and Future Work
- Bibliography / References
- Appendix

Keys to writing a scientific paper

The key to writing a scientific paper is that an independent researcher should be able to duplicate your results reported in your paper from what you convey in the paper.

- Introductions and Conclusions are extremely important
- Be consistent, e.g., references
- Diagrams are helpful
- Clean, clear, almost self-explanatory

A part of Word Treatment output of Example 1

English	Word Treatment Output	Description
spring	ฤดูใบไม้ผลิ (ry/duubajm ajphl), 2-4-3-2	the season between winter and summer
	สปริง (sapring), 1-1-2-1-1-3-1-4	an elastic device
	กระโดด (krad od), 2-1-5-5-1	to move suddenly upward or forward
	น้ำพุ (n mph), 1-1-2-1-4-1	place where water comes up naturally from the ground

TC: จอห์น (caan- John) พูด (ph ud- talk) เกี่ยวกับ (ki w k b- about) ฤดูใบไม้ผลิ (ry/duubajm ajphl - spring, the season) ที่ (th i- modifying) หัก (h g- broken) นั้น (n n- the)

TC with the new selection: จอห์น (caan- John) พูด (ph ud- talk) เกี่ยวกับ (ki w k b- about) สปริง (sapring- the device) ที่ (th i- modifying) หัก (h g- broken) นั้น (n n- the)

Repaired TC: จอห์น (caan- John) พูด (ph ud- talk) เกี่ยวกับ (ki w k b- about) สปริง (sapring- the device) อัน (?an- classifier) ที่ (th i- modifying) หัก (h g- broken) นั้น (n n- the)

The repair of Example 1

