Automatic detection of meaningful patterns in text

Mick O'Donnell
Universidad Autónoma de Madrid

Contents

1. Meaningful patterns in text
2. Automatic analysis of Genres and Registers:
   a) Automatic description of Genres/Registers
   b) Automatic Classification of texts
3. Automating Analysis of Ideology?
   - Critical Discourse Analysis
4. Conclusions

1. Meaningful Patterns in Text

- Many studies of meaning focus on meaning in isolated units/features:
  - We choose a word (or combination of words) to express a meaning
    - Should I say "*ship*", "*boat*" OR "*vessel*"?
  - We choose a particular syntactic structure to express our meaning.
    - *I spilt my coffee* OR *I have split my coffee*.

1. Meaningful Patterns in Text

- However, often meanings are not carried by individual choices, but rather by **patterns of choices within the text**:
  - We speak/write appropriately to our conception of the social context of the text, and thus the patterns of choices within our text encode our conception of this social context
  - Our patterns of choices can also reflect our ideologies: how we think about certain ideas, events, participants, etc.

1. Meaningful Patterns in Text: Register

- Register

"*The semiotic structure of a given situation type, its particular pattern of field, tenor and mode, can be thought of as resonating in the semantic system and so activating particular networks of semantic options*"

"This process specifies a range of meaning potential, or register: the semantic configuration that is typically associated with the situation type in question."

(Halliday 1978 *Language as Social Semiotic*, p123)

1. Meaningful Patterns in Text: Genre

- Genre: a way of getting things done

- Martin and Rose:

  "As children, we learn to recognize and distinguish the typical genres of our culture, by attending to consistent patterns of meaning as we interact with others in various situations."

  "such complex meanings fall into consistent patterns that make it possible for us to recognize and predict how each genre is likely to unfold"

  (J.R. Martin and D. Rose 2003 *Working with Discourse*, p7)

1. Meaningful Patterns in Text: Ideology

- Ideology:

  "Group or class 'consciousness' ... which underlies the socioeconomic, political and cultural practices of group members in such a way that their (group or class) interests are optimally realised." (van Dijk, Discourse & Power, 2008)

  "A set of beliefs, or entire belief system, through which a group or culture view the world"
  (Delin, The Language of Everyday Life, 2000)

1. Meaningful Patterns in Text: Human recognition

- Thus, texts contain patterns of meaning selections which express the social context assumed by the speaker/writer, their ideology, and the genre used to deliver their message.
- These “meaningful patterns” are not obvious on the surface, but are typically unconsciously recognised by the reader/listener:
  - The patterns of lexical choice provide a general idea about what is being discussed (Field), and thus facilitate the interpretation of subsequent messages.
  - The indications of the Genre in the text allow the prediction of the kinds of meanings that will be delivered in various stages of the text.
  - Indications of the Mode allow the listener to detect disparities between projected and actual mode (e.g., written speech, etc.)
  - Other indications provide the reader with an awareness of the writer’s assumptions of Tenor (distance, power relations, etc.), ideologies and attitudes.

1. Meaningful Patterns in Text: Automatic recognition

- If humans can recognise these patterns at least intuitively, why would we want to teach computers to recognise the patterns?

1. Pedagogy: computers can reveal the linguistic patterns associated with each genre/register. These patterns can be taught to students as part of genre-based literacy.

2. Web 2.0: automatic classification of texts in terms of genre, register and ideology means that individuals can access relevant texts more easily
   - E.g. "Only give me left-wing news reports"

3. Etc.

1. Meaningful Patterns in Text: Automatic recognition

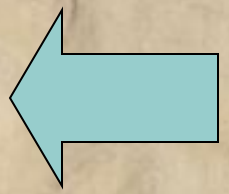
Two problems:

a) Can computers make explicit the meaningful patterns of text?
   - What lexical or grammatical patterns signal the Field/Tenor/Mode/Genre/Ideology of a text?

b) Can computers apply these patterns to automatically identify a text's social context?
   - What is the Field/Tenor/Mode/Genre/Ideology of **this** text?

Contents

1. Meaningful patterns in text
2. Automatic analysis of Genres and Registers:
   a) Automatic description of Genres/Registers
   b) Automatic Classification of texts
3. Automating Analysis of Ideology?
   - Critical Discourse Analysis
4. Conclusions



2. Automatic analysis of Genres and Register

- How do we study patterns within genres and registers?

1. Manually: human prints out texts or transcript and looks for patterns in the text.
2. By computer

Manual Text Analysis

Historic Drama ✧ Magnificent Views ✧ Nature Walks ✧ Coffee House ✧ Gift Shop

Enjoy a revolutionary view of Scotland

It's well nigh impossible to drive around Stirling without seeing the Wallace Monument. This 220ft tower dominates the surrounding plain. Take the 246 steps to the top, and you'll enjoy spectacular views.

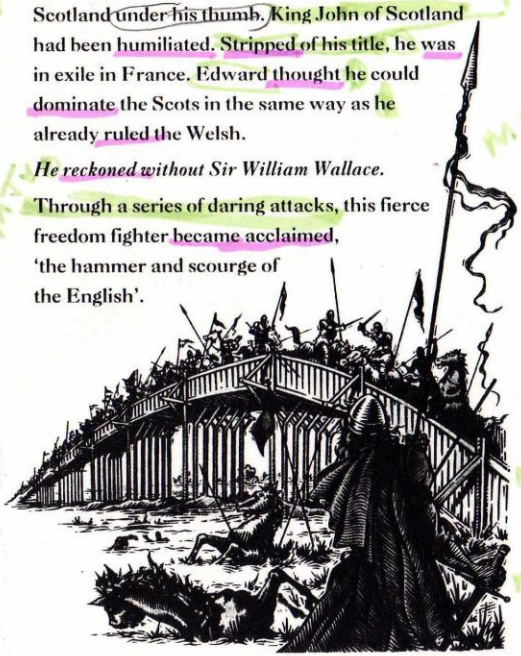
You'll also find much more to appreciate, just as revealing, on the way up.

Experience Scotland's first great battle for independence

In 1296 Edward I of England believed that he had Scotland under his thumb. King John of Scotland had been humiliated. Stripped of his title, he was in exile in France. Edward thought he could dominate the Scots in the same way as he already ruled the Welsh.

He reckoned without Sir William Wallace.

Through a series of daring attacks, this fierce freedom fighter became acclaimed, 'the hammer and scourge of the English'.



The invaders had killed his wife and brother; some revenge was gained by slaying the English sheriff of Lanark.

Marshalling a well disciplined national fighting force, Wallace became recognised as the Guardian of Scotland. With his great two-handed sword and loyal followers he cut swathes of resistance throughout the country, culminating in the siege of the English garrison at Dundee in 1297.

Advanced warfare

Edward was enraged. To deal with the rebels, he sent a massive army north: 10,000 infantrymen and 500 cavalry. The finest fighting force in Europe was armed with the most advanced weapons of the time: longbows.

It got as far as Stirling Bridge.

Wallace attacked as the army was divided by the River Forth. Over 100 English knights and 5,000 infantrymen died that day; the rest fled in disarray.

Eventually he paid a heavy price for his convictions. In 1298 his army was heavily outnumbered at the Battle of Falkirk, and destroyed. After years in hiding, he was captured and sent to trial in London. He was hung till semi-conscious, disembowelled while still alive and his body cut into quarters and displayed at Newcastle, Berwick, Stirling and Perth.

Wallace left a legacy: a belief that inspired the whole of Scotland... In 1314 King Robert the Bruce led the Scots to full nationhood at the Battle of Bannockburn.

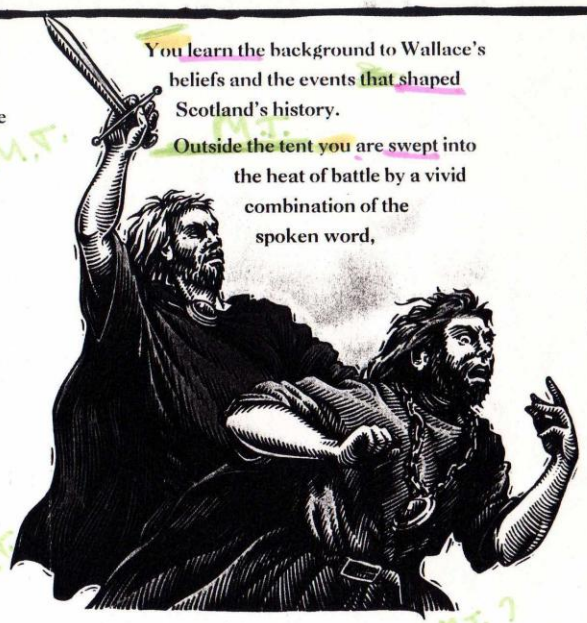
Step into the battle tent

On the first floor of the Wallace Monument you get to understand Wallace's battle tactics at Stirling Bridge and feel some of the terror that engulfed the invaders.

In the battle tent a dramatised reconstruction and a talking head of William Wallace take you through those times.

You learn the background to Wallace's beliefs and the events that shaped Scotland's history.

Outside the tent you are swept into the heat of battle by a vivid combination of the spoken word,



powerful images and evocative music. How did Wallace, a guerrilla fighter, outwit such a powerful force? You don't just learn it, you live it.

You also see Wallace's mighty two-handed broadsword. It's easy to imagine the force of it!

See far and wide

The third floor of the monument gives you a 360 degree diorama of the surrounding countryside. Through an imaginative words-and-pictures presentation, you learn about it's history. And on the viewing platform above, you enjoy the real thing: one of the most awe-inspiring views in Scotland, reaching as far as the Forth Bridges to the east and Ben Lomond to the west.

2. Automatic analysis of Genres and Register

- How do we study patterns within genres and registers?

1. Manually
2. **By computer**:
    1. Common words (keyword analysis, e.g. Wordsmith)
    2. Phrasings (lexical bundles)
    3. Grammatical Profiles

2. Automatic analysis of Genres and Register: Keyword Analysis

- Keyword analysis looks at the lexis that is more frequent in one text or text-type than in a more general corpus.
  - Scott, M. 1997. PC Analysis of Key Words -- and Key Key Words. System 25 (1): 1-13.
- Common in tools such as WordSmith (Mike Scott 2004).
- List of most relevant words in a field useful for language teaching:
  - Scott, M. and Tribble, C. 2006. Textual patterns: Keyword and corpus analysis in language education, Amsterdam: Benjamins.

2. Automatic analysis of Genres and Register: Phrasings

Phrasings:

- We can also analyse a text in terms of frequently occurring multi-word units within the genre/register/ideology.
- Early work on this area by Doug Biber (Biber et al. 1999; Biber and Barbieri 2007)
- Lexical bundles: sequences of *n* words which re-occur frequently in the corpus.
- For instance, 4 word bundles in a corpus of written documents for university students:
  - *beginning of each class,*
  - *during my office hours,*
  - *over the course of,*
  - *the beginning of each,*
  - *the end of each*

2. Automatic analysis of Genres and Register: Phrasings

Biber and Barbieri 2007

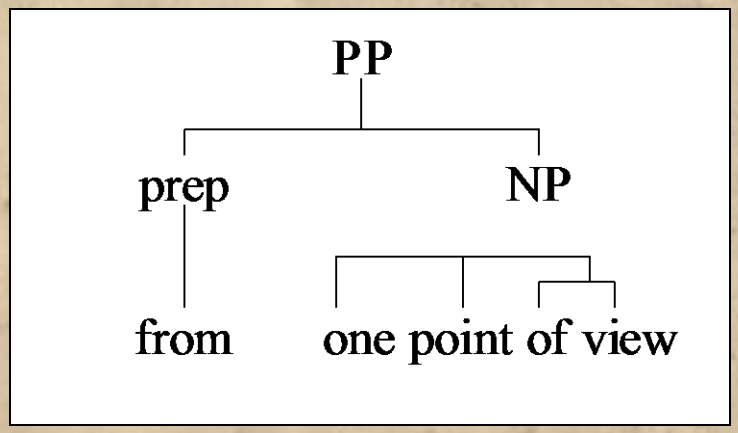
> "Lexical bundles provide a kind of pragmatic head for larger phrases and clauses, where they function as discourse frames for the expression of new information. … the lexical bundle expresses stance or textual meanings, while the remainder of the phrase/clause expresses new propositional information that has been framed by the lexical bundle."

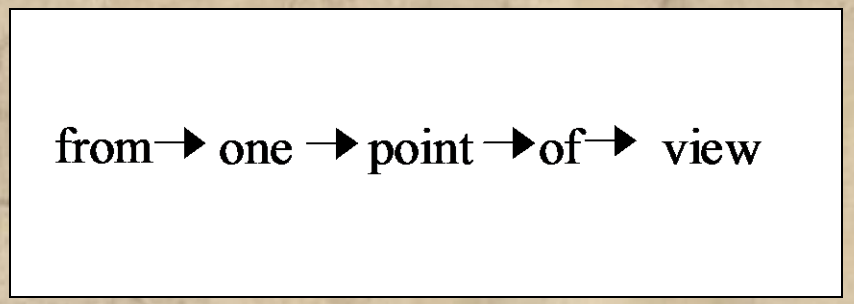
I want you to write a very brief summary of his lecture.

Hermeneutic efforts are provoked by the fact that the interweaving of …

2. Automatic analysis of Genres and Register: Phrasings

- Word sequences are a means of organising language not captured by models of syntactic structure.





- Good text conforms both to syntactic structure and to expectations of word sequence (collocation, bundles, etc.)

2. Automatic analysis of Genres and Register: Phrasings

- Lexical bundles provide very useful information about a genre/register: typical wordings which are very indexical of the type.
- We could express these meanings in other ways, but in fact, experienced writers speakers frequently re-use the phrasings of the field or genre.
- The use of these phrasings is a sign of having mastered the genre.
- For this reason, discovery of these phrasings is an important step in educating laymen into a field, or of teaching a language to non-natives.
- Related work by:
  - Michael Hoey: Lexical priming
  - Hunston and Francis: Pattern Grammar
- Software for analysis:

http://llt.msu.edu/vol10num1/review2/default.html

2. Automatic analysis of Genres and Register: Grammar

Grammar Profiling

- A genre or register can also be explored in terms of the grammatical patterns that are typical of the genre.
- In early days, such patterns were found by visually scanning texts of the genre/register.
- In later days, register studies use an electronic corpus, automatically or manually parsed.
- Biber's (1988) study famous in this regard.
- With recent advances in technology, these studies have become far easier.
- Next in this talk, I will show how easy it is

2. Automatic analysis of Genres and Register: Grammar

Grammar Profiling with UAM CorpusTool

The following study was started 2 days ago, except for some of the data collection

Texts were collected for 4 fields and 4 “text types”

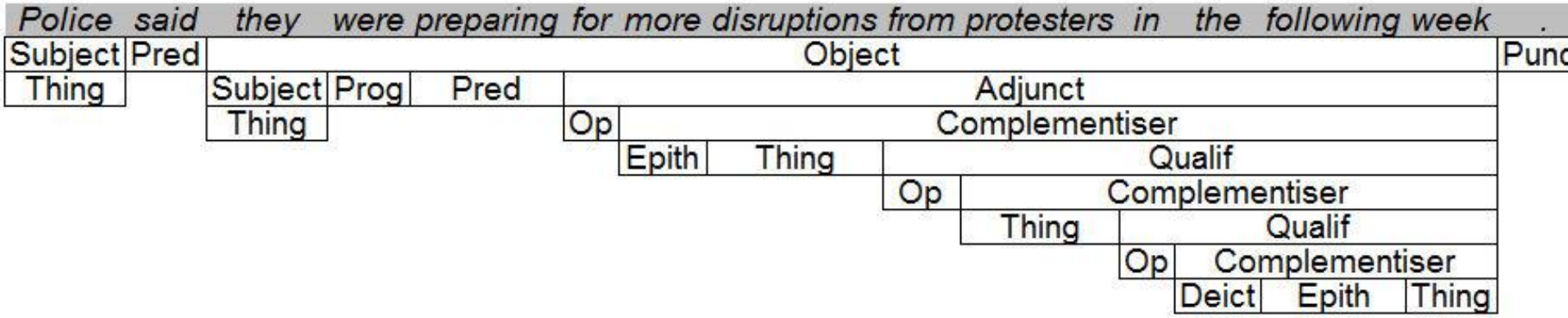
- 156 texts with 75,000 words
- Fields: medicine, military, finance, crime
- Text types:
  - Short fiction,
  - Academic abstracts,
  - Front page news (FPN)
  - Editorials

	Fiction	Abstract	FPN	editorial
medicine	8	43	3	1
military	13	8	11	8
finance	0	12	11	5

2. Automatic analysis of Genres and Register: Grammar

Grammar Profiling with UAM CorpusTool

- Texts loaded into UAM CorpusTool
- Added a “layer” of analysis to store metadata for each file (field, text-type)
- Specified that each file should be automatically parsed
  - Stanford Parser (Klein and Manning 2003) - free download from the web
  - Stanford parse converted into a more functional analysis:



2. Automatic analysis of Genres and Register: Grammar

Grammar Profiling with UAM CorpusTool

- CorpusTool then used to explore:
  - Grammatical profile of each Field and Text Type
  - Lexical Density of each Field/Text-Type
  - Keywords and Phrases

- (DEMO)

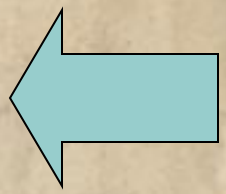
2. Automatic CLASSIFICATION of Genres and Register

Automatic Classification of Genres and Registers

- Lots of tools available for automatic text classification
- In earlier work (presented 2003), I showed how just the words in a text could be used to classify texts in relation to their field and text-type.
- Using the same Fields and Text-Types as in the study presented here:
- A test set of 40 documents:
  - 92% classified correctly by Field
  - 87% classified correctly by text-type (all mistakes between FPN and editorial)
- Errors were mostly explained by the relative small size of the corpus.
- With the addition of grammatical profiles, these results could be improved.

Contents

1. Meaningful patterns in text
2. Automatic analysis of Genres and Registers:
   a) Automatic description of Genres/Registers
   b) Automatic Classification of texts
3. Automating Analysis of Ideology?
   - Critical Discourse Analysis
4. Conclusions



3. Automating Analysis of Ideology?

- Ideology in text can be studies in text through a variety of tools.
- Two such are:
  - Critical Discourse Analysis (Kress, Fowler, Hodge, Fairclough, etc.)
  - Appraisal Analysis (Martin and White 2005)
- The question is: Can these approached be automated to work by machine?

3. Automating Analysis of Ideology? CDA

Automating Critical Discourse Theory (CDA)

- CDA has been used for 30 years to discover indications of the writer's ideology through pointing out tokens bearing ideological significance. E.g.
  - Hiding of Agency:
    - Passivisation: *Bagdad bombed yesterday.*
    - Nominalisation*: The bombing of Bagdad*
    - Use of intransitive verbs*: 35 Iraqis died* (not "killed")
  - Presuppositions:
    - *He is now an upstanding member of society*.

NOW -> HE wasn't before.

Bombing of Baghdad market kills 15

Globe & Mail March 26, 2003

The invasion of Iraq claimed its first significant civilian casualties Wednesday when a pair of massive explosions rocked a busy Baghdad marketplace, leaving charred bodies and mangled cars littering the streets.

At least 15 were killed and enraged local residents told a BBC correspondent that the death toll was in the "dozens." Another BBC reporter who visited the scene described it as "a very apocalyptic site."

Reuters News Agency reported that crowds of enraged Iraqis carried bodies away, chanting: "There is no god but Allah" and "We will sacrifice our blood and souls for you, Saddam!"

Television images showed fire engines and ambulances racing to the area as fires blazed in shattered buildings.

The U.S. military said that it was investigating the incident but that there was no reason to assume that coalition forces were responsible.

"We don't know that those were ours," U.S. Brigadier-General Vince Brooks told reporters in Qatar. "we can't say that we had anything to do with that at this point. Once we have more information, we'll be on the record."

One resident said a pair of missiles hit the busy street, which is lined by ground-floor-level shops underneath residential apartment blocks, at around 11:30 a.m. He said he believed as many as 27 people had been killed in the attack.

3. Automating Analysis of Ideology? CDA

- The question is: Can techniques be automated?
- Fowler 1991: *There is no constant relationship between linguistic structure and its semiotic significance.*
- Example: agentless passive could be used:

1. Because the agent is not relevant
   - *This book was first published in 1860*
2. To hide agency:
   - *3 protesters were killed during the protest*

- So, we cannot tell infer ideological stance just by the presence of this syntactic structure.

3. Automating Analysis of Ideology? CDA

- But in other cases? Possible study:

1. Assumption: your allies “state” information, your enemies “claim”
2. Method:
   i. Acquire list of republican and democrat politicians
   ii. Download large number of news articles from papers from known political bias.
   iii. Use reference resolution software to locate all mentions of these politicians.
   iv. Count the times each one states, claims, etc.
   v. Test whether democrat papers make more definite statements in democratic-oriented papers. And visa versa.
   vi. If so, apply same technique to new articles to classify them for political bias.

3. Automating Analysis of Ideology? CDA

- More useful: computers as tools to locate potential tokens of ideological significance:
  - All passives, especially those without agents
  - Intransitive verbs where transitive ones possible (die/kill)
  - Nominalisation of processes (the bombing)
- Then, human coder marks each as significant or not.
- Computer is a WORKBENCH for ideological investigation, not the sole agent.

4. Conclusions

1. Some meaningful patterns in text can be identified by the computer
   - The words and phrasings of registers and genres
   - The grammatical patterns that realise them.
2. Such studies are far easier now than ever before.
3. These patterns can be used with reasonable accuracy to identify the social context of new texts.
4. Ideological patterns are harder to extract from texts.
5. Basically, software can make such studies easier to perform.

5. References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber D. and F. Barbieri 2007. "Lexical bundles in university spoken and written registers". English for Specific Purposes 26, 263-286.

Hardt-Mautner, G. 2005 'Only Connect. Critical Discourse Analysis and Corpus Linguistics. ??

Scott M. 2004. WordSmith Tools version 4. Oxford: Oxford University Press.