Word Meaning and Corpus Analysis

Barcelona, May 2007

Patrick Hanks
Masaryk University, Brno
hanks@fi.muni.cz

Course outline: main themes

- Analysing corpus data for words in use
  - start with the words, not the syntax
- Building the dictionary of the future
- Mapping meaning onto use
- The theory of norms and exploitations
- Meaning and metaphor

Agenda for Monday

- Background; foundations; terminology
- Halliday and Sinclair
- Corpus Pattern Analysis: introduction

Discussion Points (1)

- What is a corpus?
- Is the Word Wide Web a corpus?

Definition of "corpus"

- A corpus is a large collection of texts in electronic form for tagging and analysis
  - part of speech tagging (word classes)
  - semantic tagging (?!)
  - parsing (??!!)
  - anaphora resolution (???!!!)
- **manually, or automatically?**
  - **PROBLEMS:**
  - **interannotator (dis)agreement (if manual)**
  - **error identification (if automatic)**

Short History of Corpora

- Brown and LOB (1960s, 1970s)
  - 1 million words each
- COBUILD (1980s)
  - 20 million words
- British National Corpus (1990s)
  - 100 million words
- Very large corpora (2000s)
  - billions of words
  - **easy to build**
  - **now build specialized domain corpora**

Size and Stability

- "More data is better data"
  - Jelinek and Mercer (at IBM), c. 1988
- "Balanced" vs. "opportunistic" corpora
  - Is a balanced corpus possible?
  - Defining the population to be sampled
  - No definition of all English text types is possible
- The WWW is not a corpus
  - because it is constantly changing
  - stability is needed, to compare like with like

Discussion points (2)

- What is a word?

Meanings of 'word'

- **Token:** an occurrence
  - "the cat sat on the mat" -- 6 tokens
- **Type:** a form
  - "the cat sat on the mat" -- 5 types
- **Lemma:** a group of forms
  - *sit, sits, sitting, sat* -- 1 lemma
- **Lexical item:** e.g.
  - *hair, keep; fire; fire engine; keep your hair on; -ing*
  - What about *gas fire, forest fire, wood fire*?
    - **Lexical items or syntactic constructs?**
    - **Linguistic categories have fuzzy boundaries**

Halliday, Sinclair, and the Lexicon

Patrick Hanks
Masaryk University, Brno

Michael Halliday: some major publications

- M. A. K. Halliday (1966): 'Lexis as a Linguistic Level'.
- M. A. K. Halliday (1969): 'Systemic Grammar' in *La Grammatica, La Lessicologia.* Società di Linguistica Italiana.
- M. A. K. Halliday and R. Husain (1976): *Cohesion in English.*

Halliday's teachers

- The linguist J. R. Firth ("you shall know a word by the company it keeps")
- The anthropologist Bronislaw Malinowski (language as an instrument of social cohesion, "phatic intercourse")

First steps in assessing what is going on in a text:

Three aspects of text:

Experiential (past, present, or possible).

Interpersonal: e.g. greetings; persuasion; performatives; the pronouns *I/you,* …

Intertextual: e.g. chains of "ties" such as anaphoric pronouns; given (*the dog*) vs. new (*a dog*); discourse organizers such as *anyway, however; …*

Cohesion

- Texts do not consist of random collections of sentences. They are coherent.
- The function of some words is to make the text cohere. Examples include:
  - Pronouns: *he/him, she/her, it, they/them*
    - Important for corpus pattern analysis
  - Some determiners: *the (*but not *a), this, these, their, …*
  - Discourse organizers: *In the first place, however, On the other hand, Against this, …*
  - Eugene Winter's *"Vocabulary 3"*

Clause Roles: SPOCA

- Subject
- Predicator: the verbal group
  - Phased predicators: *she wanted to go; she wanted him to go.*
- Object [0, 1, or 2]
  - *She gave* ***him a book.*** (2 objects)
  - *He fainted.* (0 object)
- Complement
  - Subject complement: *she is* ***happy****; he is* ***a clown****.*
  - Object complement: *she made him* ***happy****;*
    - *They appointed him* ***director of the CIA****.*
- Adverbial [Adjunct]:
  - *She gave a book* ***to him;***
  - *They treated John* ***with respect (respectfully)****;*
  - *The doctors treated James* ***with antibiotics (homeopathically);***
  - She baked a cake ***in the kitchen | with glee | yesterday | in the oven | at 4 o'clock | for Peter's birthday | …***

Rank, Exponence, Delicacy

The Rank scale (different levels of delicacy):

- Discourse
- Paragraph (in written text) / turn (in conversation)
- Sentence
- Clause
- Phrase ("Group")
- Word
- Morpheme

Rank shift

- A unit may be an exponent of an element in the rank next below it.
- Rank shifted clause (functioning as a group):
  - {the house {that Jack built}}
- Rank shifted group (into a larger group):
  - {the house {on the hill}}
- Rank shifted nominal group (functioning as a word -- e.g. a possessive determiner):
  - My dog, Fred's dog, {my {aunt's}} dog, {my {uncle's {wife's}}} dog, …

Lexical relations

- A powerful argument/a strong argument
- A powerful car/strong tea
- *a strong car/*powerful tea
  - "*strong* and *powerful* are members of a class that enters into a certain structural relation with a class of which *argument* is a member [and *tea* and *car*] ."
  - Church, Hanks, Gale, and Hindle (1990) show that collocates of ***strong*** are typically intrinsic (e.g. *strong defence*), whereas collocates of ***powerful*** are typically extrinsic (e.g. *powerful enemies*).

Patterns "reappear" in different syntactic contexts

- a strong argument …
- He argued strongly …
- … the strength of his argument
- His argument was strengthened by …

Lexis and structure

- "In place of the highly abstract relation of structure, … lexis seems to require the recognition merely of linear co-occurrence together with some measure of linear proximity."
- "In place of 'system', which lends itself to a deterministic model, lexis requires the open-ended 'set', assignment to which is best regarded as probabilistic."

Lexis and Grammar

- "Collocation and lexical set are mutually defining, as are structure and system: the set is the grouping of members with like privileges of occurrence in collocation."
- "In lexis we are concerned with a very simple set of relations into which enter a large number of items, … whereas in grammar we are concerned with very complex and variable relations in which the primary differentiation is among the relations themselves."

Lexis and Grammar (2)

- "It is essential also to examine collocational patterns in their grammatical environments and to compare the descriptions given by the two methods, lexical and lexicogrammatical."

Lexical sets

- The criterion for the assignment of items to sets is collocational.
- Halliday's (1966) prediction, collocates of ***sun***: *bright, hot, shine, light, lie, come out*
- Word Sketch (BNC statistically most significant collocates): *shine, microsystem, terrace, setting, moon, midday, rising, hot, ray, blazing, afternoon, morning, evening, … warm, bright, …*

A project for the future (1966, 2007, … when?)

- "A thesaurus of English based on formal criteria, giving collocationally defined lexical sets with citations to indicate the defining environments, would be a valuable complement to Roget's brilliant work of intuitive semantic classification, in which lexical items are arranged 'according to the *ideas* which they express'." - Halliday 1966

Collocations

- "You shall know a word by the company it keeps" -- J. R. Firth
  - **Collocations:** co-occurrences within a 4- or 5-word span either side of a node word
  - **Colligations:** co-occurrences in a syntactic relationship with the node word
- Sinclair showed how collocations work and how they affect meaning.

Sinclair: idiom and openness

- **open-choice principle:**
  - "a way of seeing language as the result of a very large number of complex choices. At each point where a unit is complete (a word or a phrase or a clause), a large range of choices opens up and the only restraint is grammaticalness"
- the **idiom principle:**
  - "Many choices within language have little or nothing to do with the world outside. … a language user has available to him or her a large number of semi pre-constructed phrases that constitute single choices."

Statistically significant collocations

- "of the" is a frequent collocation in English
  - but not very interesting.
- How to find interesting collocations like "doctors, nurses, treat, injury, health, ..."?
- Church and Hanks (1990): **mutual information** (MI)
- **t-score**

MI and t-score

- MI favours rare words: e.g.
  - "Tallulah + Bankhead" *but also*
  - "doctors, nurses"; "bread, butter"
  - MI underlies the Sketch Engine
- t-score favours function words, e.g.
- "swallow something up", "refrain from smoking"

Sinclair: some major publications

- J. M. Sinclair. 1966. 'Beginning the Study of Lexis'.
- 1970 (2004). *The OSTI Report*.
- 1991. *Corpus, Concordance, Collocation.*
- 2003. *Reading Concordances.*
- 2004. *Trust the text: Language, Corpus, and Discourse.*
- (with Anna Mauranen) 2006. *Linear Unit Grammar.*

Sinclair's OSTI report

- The nature of collocation and lexical patterning;
- The nature of the lexical item (including "multiword items"—e.g. *red herring*);
- Relationship between grammar and lexis;
- The Zipfian distribution of word frequencies;
- Differences between spoken and written language.

Evidence and Intuition

- "One does not study botany by making artificial flowers."
- "I'm interested in explaining what does occur, not what might occur." -- Sinclair
  - Hanks's version: Don't ask, "Can you say X?"
  - Ask instead, "Is it normal to say X?"

Social salience; Cognitive salience

- **Social salience:** how words are actually used
- **Cognitive salience:** how we think words are used.
- Example:
  - What's the most common use of ***total*** as a verb?
- People report cognitively salient uses when asked about "What is common?"

Sinclair's "blue jeans principle"

- The semantic lightness of frequent words:
- The more you use them (and wash them), the more the colour washes out.