

# The Generative Lexicon (GL) meets Corpus Pattern Analysis (CPA)

Patrick Hanks

Institute of Formal and Applied Linguistics,  
Charles University in Prague, Czech Republic

\*\*\*

# Where I'm coming from

- British lexicography (Collins, Oxford)
- The Firthian tradition in linguistics (Sinclair, Cobuild)
- 25 years of corpus analysis (CPA)
- Analysing data, asking what sort of theory accounts for observable patterns, how to map meaning onto use, ...
- Painful discoveries:
  - Patterns of linguistic behaviour are everywhere in corpora
  - Meanings can be mapped onto patterns
  - All too often, speculative linguistic theory (SLT) doesn't match the evidence, or is not accurately focused
- GL provides an apparatus for CPA
- CPA provides empirical support for (some) GL

# Empirical Recognition of Patterns

- When you first open a concordance, **patterns** start leaping out at you.
  - Collocations make patterns: one word goes with another
  - To see how words make meanings, we need to analyse collocations
- The more you look, the more patterns you see.

BUT

- When you try to formalize the patterns, you start to see more and more exceptions.
- The boundaries are fuzzy and there are many outlying cases.

# The linguistic 'double-helix' hypothesis

- A language is a system of rule-governed behaviour.
- Not one, but TWO (interlinked) sets of rules:
  1. Rules governing the normal uses of words to make meanings
  2. Rules governing the exploitation of norms

# Exploitations

- People exploit the rules of normal usage for various purposes:
- For **economy** and **speed**:
  - Conversation is quick
  - Listeners (and readers) get bored easily
  - Words that are ‘obvious’ can sometimes be omitted
- To **say new things** (reporting discoveries, registering patents, ...)
- To **say old things in new ways**
- For rhetoric, humour, poetry, politics ...

# Lexicon and prototypes

- Each word is typically used in one or more patterns of usage (valency + collocations)
- Each pattern is associated with a meaning:
  - a meaning is a set of prototypical beliefs
  - In CPA, meanings are expressed as ‘anchored implicatures’.
  - few patterns are associated with more than one meaning.
- Corpus data enables us to discover the patterns that are associated with each word.

# What is a pattern?

- The verb is the pivot of the clause.
- A pattern is a statement of the clause structure (valency) associated with a meaning of a verb,
  - **together with typical semantic values of each argument**, realized by salient collocates
- Different semantic values of arguments activate different meanings of each verb.

# Pattern are contrastive

*fire*, verb

1. [[Human]] fire [[Firearm]] (at [[Phys Obj = Target]])
  2. [[Human]] fire [[Projectile]] (from [[Firearm]]) (at [[Phys Obj = Target]])
  3. [[Human 1]] fire [[Human 2]]
  4. [[Anything]] fire [[Human]] {with **enthusiasm**}
  5. [[Human]] fire [NO OBJ] .... (= 1 or 2, not 3 or 4)
- Etc.

# Types and Qualia in CPA

- The apparatus needed for analysing nouns is different from that needed for verbs
  - Plug and socket
- Verbs need event typing and argument structure
- Nouns need qualia
  - What sort of thing is it?
  - What's it for?
  - What properties does it have?

AND

- Is it good or bad (and for whom)?

# Each argument of each verb is a complex lcp

- [[Event | Human]] calm [[Animate]]
  - calm a hysterical patient
  - calm the horses
  - But can you *\*calm a cockroach?*
    - Not part of the lcp for “calm [[Animate]]” – not a norm
  - Calm { [[POSDET] {nerves | anxiety} [= properties of [[Animate]] ]
  - Calm a riot [= behaviour of [[Animate]] ]
  - Calm the market [[= Location = Activity in Location = Human Group Acting in Location]]

# Semantic types and semantic roles

- **sentence**, v.
- PATTERN: [[Human 1 = Judge]] sentence [[Human 2 = Convicted Criminal]] to [[{Time Period | Event} = Punishment]]
- IMPLICATURE: [[Human 1]]
- SECONDARY IMPLICATURE: [[Time Period]] is a jail sentence
- EXAMPLE: *Mr Woods sentenced Bailey to 7 years.*

Note that the implicature is “anchored” to the pattern.

# Semantic Types and Ontology

- Items in double square brackets are **semantic types**.
- Semantic types are being gathered together into a shallow ontology.
  - (This is work in progress in the current CPA project)
  - Preliminary outline in Pustejovsky, Rumshisky, and Hanks 2004
- Each type in the ontology will (eventually) be populated with a set of lexical items on the basis of what's in the corpus under each relevant pattern.

# Shimmering lexical sets

- Lexical sets are not stable – not „all and only”.
- Example from Hanks and Jezek (2008):
  - [[Human]] attend [[Event]]
  - [[Event]] = *meeting, wedding, funeral*, etc.
  - But not all events: not *thunderstorm, suicide*.
  - and not only events: *attend school, attend a clinic*
- Contrast with another pattern for ***attend***:
  - [[Human 1]] attend [[Human 2 = High Status]]

# Meanings and boundaries

- Boundaries of all linguistic and lexical categories are fuzzy.
  - There are many borderline cases.
- Instead of fussing about boundaries, we should focus instead on identifying prototypes
- Then we can decide what goes with what
  - Many decision will be obvious.
  - Some decisions – especially about boundary cases – will be arbitrary.

# The Idiom Principle (Sinclair)

- In word use, there is tension between the „terminological tendency” and the „phraseological tendency”:
  - The **terminological tendency**: the tendency for words to have meaning in isolation
  - The **phraseological tendency**: the tendency for the meaning of a word to be activated by the context in which it is used.

# Current work in progress

- Hanks (forthcoming): *Analyzing the Lexicon: Norms and Exploitations*. MIT Press
  - A corpus-driven, lexically based theory of meaning in language
- Linked to PDEV (*A Pattern Dictionary of English Verbs*) by CPA (*Corpus Pattern Analysis*)
  - A basic infrastructure resource
  - 468 verbs analyzed and released, freely available
  - <http://nlp.fi.muni.cz/projects/cpa>
  - Experiments with automating the analytical procedure and applying the results for NLP (IR, MT, ...) and language teaching (lexical syllabus design)
  - Building a shallow ontology is in progress

# Thanks

- The late John Sinclair & colleagues (Cobuild project)
- Bob Taylor, Marie-Claire van Leunen & the late Digital Equipment Corporation Systems Research Center in Palo Alto (Hector project)
- James Pustejovsky, Anna Rumshisky, & Brandeis U.
- Masaryk U., Brno & Karel Pala, Pavel Rychly, and Adam Rambousek
- Institute of Formal and Applied Linguistics, Charles U., Prague, & Jan Hajic, Martin Holub
- Various Czech agencies for funding
- You, for listening