Corpus Pattern Analysis and Lexicography

Patrick Hanks
Research Institute of Information and Language Processing,
University of Wolverhampton

University of the West of England, Bristol
Theme of the talk

- *Question:* What is meaning?
- *Other questions:* How does meaning work?
- How does language work?
- How should monolingual dictionaries set about explaining meaning?
• The market for dictionaries as printed books has collapsed.

• Dictionaries now are products for hand-held devices and on-line resources:
  – Very variable in quality
  – Most of them list many senses for most words.
  – None of them tell the user how to distinguish one sense of a word from another.

• New types of dictionaries are needed.
  – Only the big electronics companies (such as Microsoft, Google, Facebook, etc.) can afford the level of investment required.
  – But they are not doing it.
  – Why not?
Background (2)

• Linguistics as an academic subject is in chaos.
  – In the English-speaking world, at least

• Generative linguistic theory of the kind envisioned by N. Chomsky has proved untenable.
  – It focuses on syntax, but has nothing useful to say about meaning. (It is, in short, “an impoverished and thoroughly inadequate conception of language”).
  – Incompatible with corpus data.

• New theoretical approaches are needed.

• Tenable contributions to theory have emerged from:
  – Philosophy of language (Wittgenstein, Austin, Grice, Putnam)
  – Psychology and anthropology (Rosch, Lakoff, Langacker)
  – Lit. crit. (I. A. Richards)
  – Russia (Mel’cuk)
  – Lexicography and Function (Sinclair; Halliday, Firth)
The source of the trouble

- In the past, brilliant thinkers (e.g. Wilkins, Leibniz, Russell, Wierzbicka, Chomsky, and many others) confused the relationship between logic and natural language.

- **Logic** (almost any kind of traditional logic) is crisp, neat, clean, … formalizable.
  - Crisp, neat, clean thinkers want language to be crisp, neat, and clean (wishful thinking).
  - But it isn’t.

- At its heart, natural language is fuzzy, messy, dirty. Word meanings are difficult (or impossible?) to formalize.
  - This fuzziness of meaning is a design feature of natural language, enabling people to use old words to say new things.
An empirically well-founded, corpus-driven theory of meaning in language

- The Theory of Norms and Exploitations (TNE) arises out of 30 years experience of corpus analysis and 50 years of writing and editing entries for dictionaries.
- The present talk is based on TNE.
- It aims to show how analysis of the patterns of word use that can be observed in a large corpus can lead to a better understanding of how people use words to make meanings, better teaching, and better (electronic) dictionaries.
- The procedure is called CPA (Corpus Pattern Analysis).
A working hypothesis

• Much meaning is created and understood by pattern matching (subconsciously matching word uses in texts with patterns of word use that have been sorted somehow in our brains).
  – Pattern matching is going on all the time when you speak and write, or listen and read.

• Q: Professor Hanks, what are these patterns, of which you speak? -- Answer: We don’t know.
  • Q: How can we find out? -- Answer: Through corpus pattern analysis (CPA).
A nasty surprise

• I am not a linguist. I am a lexicographer. I have no prior commitment to syntax, phraseology, or anything like that. My prior commitment is to finding out about meaning.
• After 20 years as a lexicographer and editing two major dictionaries, I came to a surprising conclusion: words don’t have meanings.
• So had I been wasting my time all those years?
• No, because words do have meaning potential.
• Meaning potentials are realized by context.
• Context is phraseology! I am driven to study phraseology by a desire to understand meaning and corpus data.
Philosophical background

• Grice (1957) posited that meanings are not just in the head
  — they are events; interactions between people:
  — between speaker (S) and hearer (H);
  — (and with displacement in time) between writer and reader
• For this to work, S and H must share a body of linguistic conventions having the same meanings.
• Grice did not specify what these conventions are.
  — He left that task to linguists and lexicographers
  — So far, we have not done a very good job.
  — We have let him down.
Lexis and grammar

• Are the conventions that underlie conversational co-operation conventions of *grammar* (syntax)?
  – No. Syntax has a role to play, but for nearly 60 years (since 1957) its role has been grossly exaggerated

• Perhaps the conventions that we rely on in conversation are *words*, with their meanings as stated in dictionaries?
  – But two decades of research in Word Sense Disambiguation (WSD) by computational linguists (using LDOCE and other dictionary resources) is now seen as a failure (Ide and Wilks 2006).
  – At least in part, this is because dictionaries don’t say enough about phraseology.

• Something else is needed.
Do Words have meaning?

Let’s think of a word:

• What’s the meaning of blow?
The meaning potential of a word

• What’s the meaning of *blow*? --

• What’s the meaning of *blow up*?
  – Destroying a building? What you do to a balloon? Lose your temper? Start to become publicly notorious? …

All of these things and more! Words are hopelessly ambiguous.

But put a word in context, and the ambiguity is reduced or eliminated.

Strictly speaking, words in isolation don’t have meaning; they have meaning potential.

Different aspects of a word’s meaning potential are activated in different contexts.
Prototypical patterns for *blow*, verb

[62 patterns for *blow*, verb] The main ones are:

- 12% the wind blows (+ direction)
- 6% the wind or an explosion blows something somewhere
- 14% a bomb or a person using explosive blows something up
- 4% the ship (house, tin, etc.) blew up
- 3% a disagreement blew up
- 4% the wind (or an explosion) blew something off
- 2% an explosion blew the windows out
Some idioms for *blow*, verb

- *Something blew the project off course* [= wrecked it]
- *This will blow the cobwebs away* [= get rid of useless old ideas]
- *He likes to blow his own trumpet* [= boast]
- *She felt she had a duty to blow the whistle on the government* [= expose wrongdoing]
- *He blew his brains out* [= killed himself]
- *She was blowing hot and cold* [= was indecisive]
- *He blew his top* [= lost his temper]
- *He blew a lot of his money on gambling* [= spent]
- *Lawrence blew my cover* [= revealed]
Is this adequate (in a bilingual English-Polish dictionary)?

• **Blow**: ozkwitnąć, rozkwitać, dąć, zadymać, zadąć, wiać, pociągnąć, pociągać, powiewać, powiąć, wionąć, podmuchiwac, przewiać, zaciągnąć, zaciągać, przechodzić, przyjść, obwiać, nadmuchiwać, nadmuchać, nawiewać, nawiać, przedmuchać, przepalić, przepalać, wysiąkać nos
The need for a new kind of resource

- Trying to account for all possible uses of a word such as *blow* is impossible.
- But accounting for the normal phraseology of a word (and building from there) is quite possible:
  - Such basic norms (patterns) can be collected in a corpus-driven dictionary of phraseology and collocations.
  - Such a dictionary does not yet exist.
  - In Wolverhampton, we are building one ([www.pdev.org](http://www.pdev.org)).
- Language learners and computer programs alike need to learn these basic patterns (“norms”), but they also need to know how the norms are exploited creatively.
Where to start?

• Start with verbs
  – and predicative adjectives

• The verb is the pivot of the clause
  – We make conversation by using clauses
  – (to make statements and ask questions)

• Nouns are different
  – nouns need a different kind of analytic mechanism
  – Bilingual dictionaries are useful in helping learners or translators find the right noun, getting the gender and spelling right, etc.
  – Adjectives are also different (not part of this talk).
Corpus Pattern Analysis (CPA)

• We need not just a dictionary with word meanings, but also:
  – an inventory of normal contexts for each word;
  – A set of rules stating how each context is either a) used normally or b) exploited to make metaphors etc.

• CPA aims, by careful analysis of data, to establish:
  – An inventory of normal phraseological conventions
  – The meaning (semantics and pragmatics) associated with each phraseological norm.

• Out of this arises a new theoretical approach – the Theory of Norms and Exploitations (TNE)
Patterns in Corpora

• When you first open a concordance for a lexical item, very often some patterns of use leap out at you.
  – Collocations make patterns: one word goes with another
  – in structures (constructions, valencies)
  – To see how words make meanings, we need to analyse contexts: valencies and collocations

• The more you look, the more patterns you see.

• BUT THEN

• When you try to formalize the patterns, you start to see more and more exceptions.

• Fuzzy boundaries between patterns

• How to make sense of the data?
(The theoretical foundations of corpus pattern analysis)

Collocations:

• “Many, if not most meanings, require the presence of more than one word for their normal realization. ...

  “Patterns of co-selection among words, which are much stronger than any description has yet allowed for, have a direct connection with meaning.”

Idiomaticity vs. Open Choice

• “The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.”

• “Tending towards open choice is what we can dub the *terminological tendency*, which is the tendency for a word to have a fixed meaning in reference to the world. ... tending towards idiomaticity is the phraseological tendency, where words tend to go together and make meanings by their combinations.”
  —Sinclair 2004. *Trust the Text*, p. 29
Semantic Types

- Understanding text meaning depends on analysis of collocations and their variants
  - Groups and sets of collocates [example from R. Moon]:
    - shivering in her shoes / quaking in his boots / shaking in their sandals
- Lexical sets are grouped according to semantic type
  - In this example, the semantic type of the nouns is [[Footwear]]
The CPA “Ontology”

A hierarchical inventory of 225 semantic types. Top types:

- [[Entity]]
  - [[Physical Object]]
    - [[Human]]
    - [[Animal]]
    - [[Artefact]]
  - [[Abstract Entity]]
    - etc.
- [[Eventuality]]
  - [[Event]]
  - [[State of Affairs]]
    - etc.

The semantic types of nouns disambiguate the verbs with which they are used.
GROUP 1: [[Human]] grasps [[Physical Object]]
It is hard to believe that bull-leapers grasped the horns and relied on the tossing movement to get them over the bull’s head.
Ursula leaned slowly back against the window-sill, one hand grasping the edge tightly while the other held her cigarette.
He grasped the handle of the door in one hand and the spoon in the other.
He reached out wildly, trying to grasp the creature, but it had moved away.
Benjamin stretched across and grasped the man’s hand.
Laura grasped Maggie by the arm.

GROUP 2: [[Human]] grasps [[Concept]]
In the end we will grasp the truth.
I was too intelligent not to be already grasping the rules of the game we played.
After fifteen minutes, Julia thought that she had grasped most of the story.
Teachers should grasp the fact that the DES can lay down details of a policy but that the Department of Employment funds it.
He could never grasp the essentials … of living in a western society.
He had not grasped that Ruby worked that day with a mere photograph.
She grasped what was happening.
GROUP 3: [[Human]] grasps [[Opportunity]]
Lawrence hoped his players would grasp the chance of cup glory.
The Prime Minister failed to grasp that opportunity.
Kylie, singing like she had never before, grasped the moment.

GROUP 4: [[Human]] grasps {nettle}:
Ian Corner, David Chell and their staff are bravely grasping the nettle of recession.
The Labour Party has failed to grasp the nettle in Monklands.
That's what the GMB need to do, to grasp the nettle, to move forward.

GROUP 5: [[Human]] grasps {at [[Physical Object]]}:
Theda had gone paler than usual, and she grasped at the bedpost for support.
The child was still crying as Alan sat down with him, but he grasped greedily for the milk.

GROUP 5a: [[Human]] grasps {at {straw}}:
Nadirpur's eyes widened. He was grasping at straws.
Patterson’s eyes flickered as if I’d given him a straw to grasp.
What a phraseological dictionary might look like

**grasp, verb**, denotes an EVENT in which someone seizes hold of something firmly and holds onto it.

1. You can grasp a physical object with your hands: *He grasped the handle of the door in one hand and the spoon in the other* | *Laura grasped Maggie by the arm.*

2. You can grasp an idea in your mind: *In the end we will grasp the truth.*

3. You can grasp an opportunity to do something: *Lawrence hoped his players would grasp the chance of cup glory* | *the Prime Minister failed to grasp that opportunity.*

4. [CONATIVE] If you grasp *at* something or grasp *for* something, you try to grasp it but may not succeed. *I grasped at the bedpost for support* | *the child grasped greedily for the milk.*

5. To *grasp the nettle* [BRITISH IDIOM] means to deal firmly and quickly with a difficult situation.

6. *grasping at straws* [IDIOM] is a variant of *clutching at straws*. See clutching at straws.
Notes on the phraseological approach

The emphasis is on explaining usage, rather than listing meanings.

• Each meaning is associated with a usage pattern, not with the word in isolation.

• Examples are chosen for typicality, not for interestingness.
  
  » Grammatical subject and grammatical object for each pattern are paradigmatic sets of lexical items sharing a common semantic type.

  » Similar, but slightly more complicated, are prepositional arguments of verbs ("adjuncts" or "adverbials" in Hallidayan terms)

• Explanations focus on normal usage, not all possible usage.

• The traditional goal of writing substitutable definitions stating necessary conditions for meaning must be abandoned.

• Entries are based on analysis of corpus evidence, not inherited from previous dictionaries.
Procedure for CPA of verbs

STEP 1: Identify statistically salient collocates of the target verb
- Using the Sketch Engine (Kilgarriff 2004)
- Organize them into constructions and patterns (first hypothesis)

STEP 2: Take a sample concordance for each word
- 250-500 examples
- from a ‘balanced’ corpus

[We use 50M words of the British National Corpus]

• Classify every line in the sample on the basis of its context
• Take further samples, if necessary to establish that a particular phraseology is conventional or if many patterns are found
• Check results against corpus-based dictionaries
• Use introspection to interpret data, but not to create data.
Classes used in CPA

• Norms (normal uses in normal contexts)
• Exploitations (e.g. coercions and ad-hoc metaphors)
• Alternations
  • e.g. [[Doctor]] treat [[Patient]] <--> [[Medicine]] treat [[Illness]]
    – Names (Midnight Storm: name of a horse, not a kind of storm)
    – Mentions (to mention a word or phrase is not to use it)
    – Errors
    – Unassignables

Every line in the sample must be classified
Alternations

There are three kinds of alternations in language:

- **Syntactic alternations**
  - e.g. *he fired the gun / the gun fired*

- **Lexical alternations**
  - e.g. *clutching at straws / grasping at straws*

- **Semantic-class alternations**
  - e.g. *treat [[Patients]] / treat (their) [[Injuries]]*
Some Syntactic Alternations

• Active / passive
• Causative / inchoative
  – *he fired the gun / the gun fired*
  – *she opened the door / the door opened*
• Unexpressed object
  – e.g. *he fired a gun at me / he fired at me / he fired*
  – (BUT NOT *she opened the door / *she opened*)
• Conative
  – e.g. *he grasped the bedpost / he grasped at the bedpost.*
• Resultative
  – e.g. *he shook his umbrella / he shook the rain off his umbrella*
Nouns

• We now move on from verbs to nouns.
• Nouns need a different kind of analytic mechanism:
  – And a different way of presenting collocations.
• Noun + verb collocations are syntagmatically fixed:
  – No problem; can be presented just like verb patterns.
• But nouns (noun-y nouns) have other statistically significant collocates, with which they are not in a stable syntagmatic relation.
  – “Noun-y nouns” are words like tree, car, money, idea, and shower [next 3 slides]
  – As opposed to nominalizations, e.g. distribution.
Phraseology of shower, n. (1)

1. A shower is a weather event: a short downpour of rain.
   – MWEs and alternates are: snow showers, wintry showers, showers of hail and sleet; a heavy shower, a light shower; April showers; scattered showers; occasional showers, the odd shower.
   – Showers sweep over or across locations
   – After a short time, a shower dies away or dies out, at which time the shower is said to be clearing
   – People get caught in a shower
   – Metaphors in science: showers of particles (nuclear physics); showers of meteorites or meteors (astronomy)

1.1 What a shower! (U.K. slang, derogatory) = what a group of useless, unattractive human beings!
Phraseology of shower, n. (2 & 3)

2. A *shower* is an *artefact* for pouring a continuous flow of water in droplets, simulating rainfall, over a person

- Typically, a shower is *provided* by an architect or house designer and *installed* by a builder, either in a *cabinet* in the *bathroom* of a house, or above the *bath*, or in a separate *shower-room*.
- An *en suite shower* is one that is installed in a room adjacent to a *bedroom*.
- When installed correctly, a shower *works*.
- Types of shower: *electric shower*, *power shower*, *gravity-fed* shower [and various trade names]
- People *switch* (or *turn*) a shower *on* in order to *use* it and *switch* (or *turn*) it *off* after use.

3. A shower is also a *location* with such an artefact fixed high up in it, so that it can pour water in a steady flow of droplets over a person, such that the person *stands* in the shower in order to *wash* his or her *hair* and/or *body*.
Phraseology of *shower*, n. (4)

4. A *shower* also denotes an *event* (involving human activity), in which a person uses a shower (2):
   - A person *takes* a shower or *has* a shower.
   - A shower may be *hot, cool, or cold.*
   - Taking a shower is *refreshing.*

If you have mastered all the phraseology on the last three slides, you are as well qualified as any native speaker to talk idiomatically in English about showers.
Exploiting Norms

- But people (including learners) don’t just want to use language normally and regularly (and boringly),
- They want to be creative, imaginative
- They want to use language in new and interesting ways.
Norms and Exploitations

• In order to understand meaning in language, it is essential to distinguish between:
  – norms (the basic shared conventions that S and H mutually rely on – including conventional metaphors), and
  – exploitations (freshly created metaphors and other tropes, unusual phrasing, etc.)

• Two different rule systems.
• The two rule systems interact.
• Grice again (1975): relevance theory
  – people also communicate by exploiting norms of linguistic behaviour, as well as by conforming to them
Regular and irregular linguistic performance

• Norms are first-order regularities of linguistic behaviour (usage)
• Alternations are second-order regularities of linguistic behaviour
• Exploitations are irregularities, deliberately chosen by a speaker or writer for rhetorical or literary effect
• Mistakes are irregularities that occur accidentally, not deliberately
Exploitations: what to ignore when writing a dictionary

- Exploitations are unusual uses of words, coined for rhetorical effect, economy of space, etc.
- Exploitations are deliberate and create new meanings.
- Exploitations are among the most interesting uses of words in a language.
- Sadly, lexicographers have a duty to ignore them.
Exploitation rule 1: ellipsis (omitting the obvious)

• I hazarded various Stuartesque destinations such as Bali and Istanbul.
  – Julian Barnes

  – In isolation, this sentence is incomprehensible.
  – But in context, the meaning is clear.
  – (The phrase “a guess at” has been omitted, “because it’s obvious”. See next slide.)
Stuart needlessly scraped a fetid plastic comb over his cranium.

‘Where are you going? You know, just in case I need to get in touch.’

‘State secret. Even Gillie doesn’t know. Just told her to take light clothes.’

He was still smirking, so I presumed that some *juvenile guessing game* was required of me. I *hazarded* various Stuart-esque destinations like Florida, Bali, Crete and Western Turkey, each of which was greeted by a smug nod of negativity. I *essayed* all the Disneylands of the world and a selection of tarmacked spice islands; I *patronised* him with Marbella, *applauded* him with Zanzibar, tried *aiming straight* with Santorini. I got nowhere.

• (Other *exploited* verb uses in this extract are in italics)
Exploitation Rule 2: Anomalous argument

- Always vacuum your moose from the snout up, and brush your pheasant with freshly baked bread, torn not sliced.
  —from The Massachusetts Journal of Taxidermy, 1986 (per Associated Press newswire)

- Can you vacuum a moose? ... Is it normal?
- “Can you say X in English? – the wrong question to ask. Ask instead, “Is it normal?”
Exploitation Rule 3: Metaphor

• Stoke Mandeville station is a little oasis; clean and bright and friendly.
• New Town Hotel -- a relaxing oasis for professional and business men.
• Driffield, which was a pleasant oasis in the East Riding of Yorkshire.
• The planned open-cast site was a pleasant oasis in a decaying industrial landscape.
• She regards her job as an oasis in a desert of coping with Harry’s illness
• … an oasis in the midst of this desert of feuding.

An oasis in English (and other European languages) is prototypically pleasant, relaxing, calm, and surrounded by barren, nasty desert. (The reality may be very different. What’s the prototype of the equivalent concept in Arabic?)
# Salient collocates for ‘oasis’ (SkE)

BNC freq for ‘oasis’: 307

<table>
<thead>
<tr>
<th>Collocate</th>
<th>Co-occurrences</th>
<th>Salience score</th>
</tr>
</thead>
<tbody>
<tr>
<td>greenery</td>
<td>3</td>
<td>8.11</td>
</tr>
<tr>
<td>serenity</td>
<td>2</td>
<td>7.53</td>
</tr>
<tr>
<td>desert</td>
<td>12</td>
<td>7.07</td>
</tr>
<tr>
<td>calm</td>
<td>7</td>
<td>7.28</td>
</tr>
<tr>
<td>lush</td>
<td>2</td>
<td>6.82</td>
</tr>
<tr>
<td>tranquillity</td>
<td>2</td>
<td>6.76</td>
</tr>
<tr>
<td>peaceful</td>
<td>3</td>
<td>5.75</td>
</tr>
<tr>
<td>welcome</td>
<td>4</td>
<td>5.68</td>
</tr>
<tr>
<td>pleasant</td>
<td>3</td>
<td>5.12</td>
</tr>
<tr>
<td>tropical</td>
<td>4</td>
<td>5.07</td>
</tr>
</tbody>
</table>
Measuring Collocations

- Collocations: “You shall know a word by the company it keeps.” – J. R. Firth.
- Patterns: “We must distinguish from the general mush of goings-on those elements which appear to be part of a patterned process.” – J. R. Firth.
- The meaning of a word in context depends to a large extent on its collocational preferences.
- Collocations in corpora can be measured. See www.sketchengine.co.uk/
Implications of all this (1)

- **Nouns** are referring expressions.
  - They have a ‘plug’ on them (just like a hair dryer).
  - Nouns represent concepts (and the world).

- **Verbs** are ‘power sockets’: you plug some nouns into slots around a verb in order to do things: make propositions, ask questions, interact socially, etc.

- **PROCEDURE**: We can solve the ‘word sense disambiguation problem’ by side-stepping it:
  - Patterns with verbs in them are unambiguous.
  - At RIILP, we are building an inventory of patterns – PDEV.
  - For any sentence from an unseen text, find the verb, find the best-match pattern, and PDEV will give you a meaning.
Implications of all this (2)

• Meanings in language are associated with words in prototypical phraseological patterns (not words in isolation).
• Meanings in text are interpreted by pattern matching – mapping bit of text onto the patterns in our heads.
  – The patterns in our heads come from ‘lexical priming’ (Hoey 2005)
  – Members of a language community share primed patterns
• Some uses match well onto patterns; these are ‘norms’
• Some uses seem surprising; these are ‘exploitations of norms’ [or mistakes].
• For each language, a corpus-driven lexical database will identify the normal phraseology associated with each word
• A set of exploitation rules is needed to explain creative usage.
A “double-helix” theory of meaning in language

- A human language is a system of rule-governed behaviour
  - But not one, monolithic rule system.
- Rather, it is two interlinked systems of rules:
  - 1) Rules governing normal usage
  - 2) Rules governing exploitation of norms.
- The two systems interact, producing new norms:
  - Today’s exploitation may be tomorrow’s norm.
Browse it for yourself

• A Pattern Dictionary of English Verbs
• Currently being created by Corpus Pattern Analysis:

  www.pdev.org.uk

– Related projects are starting for Spanish (Irene Renau Araque; Universidad Catolica de Valparaiso, Chile) and for Italian (Elisabetta Jezek; Universita degli Studi, Pavia)