Introduction to Corpus Pattern Analysis: Mapping Meaning onto Patterns of Words Use

Patrick Hanks
Masaryk University, Brno

Dictionaries before corpora

• Based on collections of citations (literary, not everyday)
• And/or based on introspection (made-up examples)
• James Murray (OED, 1878): “The editor and his assistants have to spend precious hours searching for examples of common everyday words. Thus, in the slips we have 50 examples of *abusion*, but of *abuse* not 5.”

Definitions before corpora

• attempted to state necessary conditions for the meaning of each word.
  – It was assumed (wrongly) that this would enable people to use words correctly.
  – Lexicographers and grammarians gave “idealized” representations of the language – i.e. we distorted it.
  – Hierarchical ontologies such as WordNet still operate in this “idealizing” tradition.

Definition: empirical observation, or stipulation of conditions?

Latin *definire* (“frequent in Cicero” – Lewis & Short) is based on *finitis* ‘end’, ‘boundary’: to stipulate boundaries.
• *unum hoc definio, tantam esse necessitatem virtutis* ‘this one thing I define (or stipulate) as being a necessary condition for virtue.’ – Cicero
  – clearly a stipulation, not an empirical observation about word meaning.

A crucial difference

• Scientific concepts and stipulative terminology:
  – Neat, tidy, orderly, lifeless.
  – If word meanings were governed by necessary conditions, you couldn’t use existing words to say new things.
• Word meanings:
  – Messy, chaotic, dynamic.
  – It’s the ‘looseness of fit’ that enables us to use existing words to say new things.

Definitions in dictionaries of the future

• Will associate meanings with *words in context*, not with words in isolation.
• What sort of contexts? – *Normal* contexts.
• How to determine the normal contexts in which words are used? – By corpus pattern analysis.
What are the components of a normal context? – (1) Verbs

The apparatus for corpus pattern analysis of verbs:
• Valencies (NOT “NP VP” BUT “SPOCA”).
• Semantic values for the lexical sets in each valency slot: [[Event]], [[Phys Obj]], [[Person]], [[Location]], etc.
  – Lexical sets can be populated by cluster analysis of corpora.
• Subvalency items (quantifiers, determiners, etc.):
  – ‘Something took place’ vs.
  – ‘Something took its place’.

Implicatures: taking prototypes seriously

When a pilot files a flight plan, he or she informs ground control of the intended route and obtains permission to begin flying.
…If someone files a lawsuit, they activate a procedure asking a court for justice.
When a group of people file into a room or other place, they walk in one behind the other.
(There are 14 such patterns for file, verb.)

A problem: deciding relevant context

Notice how the meaning of treat changes with context:
• Peter treated Mary.
• Peter treated Mary badly.
• Peter treated Mary with respect.
• Peter treated Mary with antibiotics.
• Peter treated Mary for her asthma.
• Peter treated Mary to a fancy dinner.
• Peter treated Mary to his views on George W. Bush.
• Peter treated the woodwork with creosote.

The CPA method

• Create a sample concordance for each word
  – 300-500 examples
  – from a ‘balanced’ corpus (i.e. general language)
    [We use the British National Corpus, 100 million words, and the Associated Press Newswire for 1991-3, 150 million words]
  – 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
• Check results against corpus-based dictionaries.
• Use introspection to interpret data, but not to create data.

In CPA, every line in the sample must be classified

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

Sample from a concordance (unsorted)

Incessant noise and bustle had abated. It seemed everyone was up after dawn the storm suddenly abated. Ruth was there waiting when
Thankfully, the storm had abated, at least for the moment, and storm outside was beginning to abate, but the sky was still ominous
Fortunately, much of the fuss has abated, but not before hundreds of
he declared the recession to be abating,
only hours before the
'soft landing' in which inflation abates but growth continues moderate
the fearful noise abated in its intensity, trailed
ability. However, when the threat abated in 1989 with a ceasefire in
bag to the ocean. The storm was abating rapidly, the evening sky
ferocity of sectarian politics abated somewhat between 1931 and
storm. By dawn the weather had abated though the sea was still angry
the dispute showed no sign of abating yesterday. Crew in
Sorted (1): [[Event = Storm]] abate [NO OBJ]
dry kit and go again. The storm abated a bit, and this is no problem in
ling. Thankfully, the storm had abated, at least for the moment, and the
sting his time until the storm abated but also endangering his life, Ge
outside was beginning to abate, but the sky was still ominously a
bag to the ocean. The storm was abating rapidly, the evening sky clearin
after dawn the storm suddenly abated. Ruth was there waiting when the h
t he wait until the rain abated. She had her way and Cobert went
by dawn the weather had abated though the sea was still angry, I
White, and the gales had abated: Yachting World had performed the
rain, which gave no sign of abating, knowing her options were limite
became a downpour that never abated all day. My only protection was
the roar of the wind abating as he drew the hatch closed behi

Sorted (2): [[Event = Problem]] abate [NO OBJ]
'soft landing' in which inflation abated but growth continues makes
fortunate, much of the fuss has abated, but not before hundreds of
the threshold. The fearful noise abated in its intensity, trailed
incessant noise and torture had abated. It seemed everyone was up
ability. However, when the threat abated in 1988 with a ceasefire is
the Intifada showed little sign of abating. It is a cliche to say that
be declared the recession to be abating, only hours before the pub
be ferciity of sectarian politics abated somewhat between 1931 and 1
been inverted and street violence abated, the ruling party stopped b
the dispute showed no signs of abating yesterday. Crews in

Sorted (3): [[Emotion = Negative]] abate [NO OBJ]
right on the table and his anxiety abated a little. This talented, if
that her initial awkwardness had abated if she had never seen s
e to some inner pressure doesn’t abate he wanted to play at the fun
aker is the fryer and my anxiety abated; he seemed disappointed and
hated at the time. When the agony abated he was prepared to laugh at
self; the pain gradually began to abate spontaneously, a great relie
ght, after the shock had begun to abate, the vision of Benedikt’s an
y plan, control it!!) The fear was abating, the trembling beginning t
his dark eyes. That fear did not abate when, briefly, he halted. For

AN EXPLOITATION OF THIS NORM:
Disapproval, his kindlier feelings abated, to be replaced by a resin
(“kindlier feelings” are normally positive, not negative.)

Part of the lexical set [[Event = Problem]] as subject of ‘abate’
From BNC: {fuss, problem, tensions, fighting, price war, hysterical
media clamp, disruption, slump, inflation, recession, the Mozart
frenzy, working-class militancy, hostility, intimidation, ferocity of
sectarian politics, diplomatic isolation, dispute, …}
From AP: {threat, crisis, fighting, hijackings, protests, tensions, anti-
Japanese fervor, violence, bloodshed, problem, crime, guerrilla attacks,
riot, meddling, shooting, artillery duels, fire-code violations, unrest,
inflationary pressures, layoffs, bloodletting, revolution, murder of
foreigners, public furor, eruptions, bad publicity, outbreak, jeering,
crime, insouciance, insouciant, problem, …}
(All these are kinds of problem.)

Part of the lexical set [[Emotion = Negative]] as subject of ‘abate’
From BNC: {anxiety, fear, emotion, rage, anger, fury, pain,
agon, feelings, …}
From AP: {rage, anger, panic, animosity, concern, …}

A domain-specific norm:
[[Person | Action]] abate [[Nuisance]]
(DOMAIN: Law. Register: Jargon)

- undertake further measures to
- be taken to
- or the local authority may
- and do whatever
- or the nuisance of bugles further
- has abated, but the decision is the dec
- that the nuisance could be
- to abate the nuisance. They have full jurisdic
- or the local authority may
- abate the nuisance and do whatever

odour nuisance
abated
the nuisance
abate
the nuisance
abate
the nuisance
abate
the nuisance
abate
the nuisance
abate
noise and bustle
abated
the threat
abated
the odour
abate
the odour
abate
the odour
abate
the odour
abate
the odour
abate
noise and bustle
abated
the threat
abated
the odour
abate
the odour
abate
the odour
abate
noise and bustle
abated
the threat
abated
the odour
abate
the odour
abate
the odour
abate
noise and bustle
abated
the threat
abated
the odour
abate
the odour
abate
the odour
abate
noise and bustle
abated
the threat
abated
Lexical sets are contrastive sets

- Different lexical sets generate different meanings.
- The lexical sets associated each sense of each verb are different.
  - It remains to be discovered whether they are ‘transferable’.
- In principle, lexical sets are open-ended.
- In practice, a lexical set may have only 1 or 2 members, e.g. take a {look | glance}.
- No certainties in word meaning; only probabilities.
  - ... but probabilities can be measured.
- This is where syntax meets semantics.

A more complicated verb: ‘take’

- 61 phrasal verb patterns, e.g. [[Person]] take [[Garment]] off
  [[Plane]] take off
  [[Human Group]] take [[Business]] over
- 105 light verb uses (with specific objects), e.g. [[Event]] take place
  [[Person]] take {photograph | photo | snaps | picture}
  [[Person]] take {the plunge}
- 18 ‘heavy verb’ uses, e.g. [[Person]] take [[PhysObj]] [[Adv[Direction]]
- 13 adverbial patterns, e.g. [[Person]] take [[TopType]] seriously
  [[Human Group]] take [[Child]] {into care}
- TOTAL: 204, and growing (but slowly)

A fine distinction: ‘take + place’

- [[Event]] take {place}: A meeting took place.
- [[Person 1]] take [[[[Person 2]]’s place]:
  - George took Bill’s place.
- [[Person]] take {[[COREF POSDET]] place}: Wilkinson took his place among the greats of the game.
- [[Person=Competitor]] take {[[ORDINAL]] place}: The Germans took first place.

Noun norms

- Norms for nouns are different in kind from norms for verbs.
- Adjectives and prepositions are more like verbs than nouns.
- A different analytical apparatus is required for nouns.
- Prototype statements for each true noun can be derived from a corpus.

What are the components of a normal context? – (2) Nouns

The apparatus for CPA (corpus pattern analysis) of nouns:

- Collocations.

Arranging collocates: storm (1)

WHAT DO STORMS DO?

- Storms blow.
- Storms rage.
- Storms lash coastlines.
- Storms batter ships and places.
- Storms hit ships and places.
- Storms ravage coastlines and other places.
Arranging collocates: storm (2)

BEGINNING OF A STORM:
• Before it begins, a storm is brewing, gathering, or impending.
• There is often a calm or a lull before a storm.
• Storms last for a certain period of time.
• Storms break.

END OF A STORM:
• Storms abate.
• Storms subside.
• Storms pass.

Arranging collocates: storm (3)

WHAT HAPPENS TO PEOPLE IN A STORM?
• People can weather, survive, or ride (out) a storm.
• Ships and people may get caught in a storm.

Arranging collocates: storm (4)

WHAT KINDS OF STORMS ARE THERE?
• There are thunder storms, electrical storms, rain storms, hail storms, snow storms, winter storms, dust storms, sand storms, tropical storms…
• Storms are violent, severe, raging, howling, terrible, disastrous, fearful, ferocious…

Arranging collocates: storm (5)

TYPICAL QUALITIES OF STORMS:
• Storms, especially snow storms, may be heavy.
• An unexpected storm is a freak storm.
• The centre of a storm is called the eye of the storm.
• A major storm is remembered as the great storm (of [[Year]]).

STORMS ARE ALSO ASSOCIATED WITH rain, wind, hurricanes, gales, and floods.

Why norms are important

These statements about abate and storm represent typical usage as well as typical meaning.
• They are empirically well founded (corpus-derived).
• This is where syntax meets semantics.

Exploitations

• People don’t just say the same thing, using the same words repeatedly.
• They also exploit norms in order to say new things, or in order to say old things in new and interesting ways.
• Exploitations include metaphor, ellipsis, word creation, and other figures of speech.
• Exploitations are a form of creativity.
Types of exploitation (1)

Dynamic metaphor:
- Dubrovnik became a mousetrap...
  - Associated Press (1991)
- A geometrical proof is a mousetrap
  - Schopenhauer

(Note: conventional metaphors are not exploitations: they are just a particular kind of norms.)

Types of exploitation (2)

Ellipsis:

I hazarded various Stuartesque destinations like Bali, Florence, Istanbul ...

(The norm is: [[Person]] hazard {guess}.)

- There are many other types of exploitation.

The biggest challenges currently facing CPA

- Finding quick, efficient, reliable, automatic or semi-automatic ways of populating the lexical sets.
- Applying machine-learning techniques
- Deciding “the right” level of generalization

How is CPA different from FrameNet?

CPA investigates syntagmatic criteria for distinguishing different meanings of polysemous words, in a “semantically shallow” way.

FrameNet:
- expresses the deep semantics of situations (frames);
- proceeds frame by frame, not word by word;
- analyses situations in terms of frame elements;
- studies meaning differences and similarities between different words in a frame;
- does not explicitly study meaning differences of polysemous words;
- does not analyse corpus data systematically, but goes fishing in corpora for examples in support of hypotheses;
- has problems grouping words into frames, and misses some;
- has no established inventory of frames;
- has no criteria for completeness of a lexical entry.

Goals of CPA

- To map meaning onto use.
- To create an inventory of semantically motivated syntagmatic patterns, so as to reduce the ‘lexical entropy’ of each word.
- To develop procedures for populating lexical sets by computational cluster analysis of text corpora.
- To collect evidence for the principles that govern the exploitations of norms, so that a typology can be developed