Discipline-specific Academic Phraseology: Corpus evidence and applications

Geraint Paul Rees

13th Teaching and Language Corpora Conference, University of Cambridge, 18 - 21 July 2018

I gratefully acknowledge support from the research project *significado léxico dependiente del contexto y polisemia, y las implicaciones para la representación lexicográfica* (referencia FFI2015-70375-P (MINECO/FEDER))
Overview of the Presentation

Context
• Current approaches to vocabulary selection
• Objectives of the study

Corpus-based Experiment
• Theoretical framework
• Methodology

Results & Analysis
• Summary of statistical analysis
• Examples of typical qualitative findings: treat, explain, accept

Discussion
• Possible pedagogical applications
Current Approaches to Vocabulary Selection

• **Two main approaches:**
  1. Discrete-item lists of general academic vocabulary


**Problematic:**
• Academic vocabulary is so wide-ranging that no single wordlist could meet the needs of all EAP students.
• Words exhibit distinct meanings in different academic disciplines.
• Word meaning is conditioned by phraseology
Current Approaches to Vocabulary Selection

- **Two main approaches:**
  2. The ‘Lexical Bundle’ approach where strings of words are selected according to frequency of occurrence

  **Example:** (Biber & Conrad, 1999; Simpson-Vlach & Ellis, 2010).

  **Problematic:**
  Although this approach accounts for collocates, there is **no consideration of semantic properties** when extracting phrases from text.
Objectives

• To examine differences in meaning and use of vocabulary between academic disciplines which might be obfuscated by the distributional approaches to vocabulary selection

This objective entails:

• Demonstrating a feasible means of phrase extraction which accounts for both semantic and syntactic concerns.
• Creating guidelines for producing, and an illustrative example of, a useful lexicographical resource for the EAP community.
Theoretical Framework: The Theory of Norms and Exploitations (TNE)

• Collocates arranged in lexical sets (group of words that share one or more semantic features) according to their collocational preference

• Mapped to syntactic structures as colligations

• Language as a double helix:
  “language consists of a constantly moving and developing double helix of rules governing linguistic behavior: normal uses and exploitations of normal use.” (Hanks, 2013:167)
Theoretical Framework: TNE

- Norms and their exploitations are discipline-specific.

- They can be calculated from corpus lines by examining contextually determined default interpretations

- A Practical application of TNE: **Corpus Pattern Analysis (CPA)** (Hanks, 2004; Hanks & Pustejovsky, 2005)

  Example KWIC line: *Neurosurgeons had responsibility for treating paralysis*
  Pattern: [[Human]] *treats* [[Disease]]
  Implicature: [[Human 1]] applies a [[Drug]] or [[Process]] to [[Human 2]] for the purpose of curing the patient`s [[Disease]].
Methodology: Materials

• The Academic Journal Article Corpus (second version) (AJAX2) – broadly representative of the language EAP users encounter and strive to produce.

• The Sketch Engine (Kilgarriff, Rychlý, Smrž, & Tugwell, 2004) – corpus query platform used in compilation and processing of corpus.

• The Pattern Dictionary of English Verbs (PDEV) (Hanks, 2001)

• ATLAS.ti – CAQDAS software used in annotation and analysis processes.
Methodology: Corpus Compilation

Representative

- Only included journals which were in the first quartile of JCI (SciMago, 2007) for ≥ 6 of 11 years sampled.
- No narrow sub-disciplines included.
- Sampled with equal periodicity over 11 years

Balanced

- Equal number of RAs per issue

Result: Three discipline-specific sub-corpora of approximately equal size.
Methodology: Corpus Compilation

The resulting corpus contains:

8,088,429 words in three discipline-specific sub-corpora comprising 880 RAs from 20 journals:

- **History**: 2,840,024 words in 264 RAs from 6 journals
- **Management Studies**: 2,641,811 words in 264 RAs from 6 journals
- **Microbiology**: 2,606,594 words in 352 RAs from 8 journals
Methodology: Sampling Verbs

1286 completed PDEV Verbs

≥20 occurrences

Judgements about meaning (Sinclair, 2005)

In ≥ 10 articles by distinct authors

Avoid over representing idiosyncratic uses

Final sample: 85 verbs

30 verbs randomly selected
Methodology: The Verb Sample

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>absorb</td>
<td>62</td>
<td>63</td>
<td>26</td>
<td>appear</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>manage</td>
<td>150</td>
<td>300</td>
<td>22</td>
</tr>
<tr>
<td>accelerate</td>
<td>44</td>
<td>43</td>
<td>45</td>
<td>apply</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>need</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>accept</td>
<td>300</td>
<td>157</td>
<td>65</td>
<td>call</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>note</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>accommodate</td>
<td>95</td>
<td>70</td>
<td>37</td>
<td>conduct</td>
<td>150</td>
<td>300</td>
<td>150</td>
<td>overcome</td>
<td>138</td>
<td>150</td>
<td>112</td>
</tr>
<tr>
<td>accompany</td>
<td>150</td>
<td>71</td>
<td>95</td>
<td>explain</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>prescribe</td>
<td>26</td>
<td>41</td>
<td>22</td>
</tr>
<tr>
<td>accomplish</td>
<td>73</td>
<td>138</td>
<td>67</td>
<td>fail</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>propose</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>account</td>
<td>111</td>
<td>300</td>
<td>100</td>
<td>follow</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>replicate</td>
<td>30</td>
<td>144</td>
<td>150</td>
</tr>
<tr>
<td>accumulate</td>
<td>42</td>
<td>111</td>
<td>145</td>
<td>handle</td>
<td>74</td>
<td>81</td>
<td>21</td>
<td>see</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>act</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>lead</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>specify</td>
<td>70</td>
<td>300</td>
<td>54</td>
</tr>
<tr>
<td>answer</td>
<td>145</td>
<td>89</td>
<td>41</td>
<td>maintain</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>treat</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>
After annotating 16,420 KWIC lines...

**Analysis**
- **Statistical Analysis**
  - Initial indication of differences in verb behaviour across disciplines
- **Qualitative Analysis**
  - (10 verbs)
- **Inter-disciplinary Comparison**
  - (30 verbs)
  - Examines the significance of differences in relative frequency of pattern types
- **Intra-disciplinary Comparison**
  - (16 verbs)
  - Controls for the ‘natural’ variation in frequency of pattern type between any texts
Summary of Statistical Analysis

Assuming a LL score indicating significance at $p < .05$ and a BIC score of $\geq 2$ indicating positive evidence against null-hypothesis:

**Interdisciplinary comparison:**

$H_0$: no diff in the freq. of a given pattern type across sub-corpora

$H_0$ rejected for $\geq$ one pattern type for all verbs except *accomplish*, *conduct* (single pattern)

**Intradisciplinary comparison:**

$H_0$: diff in the freq. of a given pattern type across sub-corpora

$H_0$ rejected for all 16 verbs tested

This indicates that in many cases verbs exhibit different prototypical meanings in different disciplines.
Statistical Analysis Example: *treat* (N = 150 in all)

**PT1:** [[Human 1]] or [[Institution 1]] or [[Animal 1]] treats [[Human 2]] or [[Animal 2]] or [[Entity]] or [[Event]] [Manner] 2xMicro

**PT2:** [[Human 1]] or [[Process]] or [[Drug]] treats [[Human 2]] or [[Animal]] or [[Disease]] or [[Injury]] 2xMicro

**PT3:** [[Human]] or [[Device]] treats [[Inanimate]] (with [[Stuff]] or by [[Process]]) MicroxHist

**PT4:** [[Human 1]] treats [[Human 2]] or [[Self]] to [[Eventuality]]

**PT5:** [[Human]] treats [[Anything]] 2xHist

**PT6:** [[Human 2]] treats with [[Human 2]]

**PT7:** [[Human]] treats [[MB entity 1]] or [[Animal]] with [[MB entity 2]] or [[Drug]] 2xMicro
Qualitative Analysis: *treat*

The split can be explained by the semantic type of the recipient of the action and the subject focus of each discipline:

**PT1:** 

```
```

*Implicature:* 

```
```

(= 81.33% in hist.; = 90.67% in man.; = 7.33% in microbiology)

**PT2:**

```
[Human 1] or [Process] or [Drug] treats [Human 2] or [Animal] or [Disease] or [Injury]
```

*Implicature:* 

```
[Human 1 = Health Professional] applies a [Drug] or [Process = Medical] to [Human 2 = Patient] for the purpose of curing the patient`s [Disease] or [Injury]
```

(= 6% in hist.; = 5% in man.; = 39.33% in microbiology)
Qualitative Analysis: *specify* \( (N = 70 \text{ hist}; 300 \text{ man}; 54 \text{ micro}) \)

A discipline-specific norm?

PT3: \([[\text{MB Entity 1}]] \text{ specifies } [[\text{MB Entity 2}]] \) \( (= 53\% \text{ in micro.}) \)

*Implicature*: \([[\text{MB Entity 1}]] \text{ contains the instructions necessary for the creation of } [[\text{MB Entity 2}]] \)

Or an exploitation of PT1?

PT1: \([[\text{Human}]] \text{ or } [[\text{Institution}]] \text{ or } [[\text{Concept}]] \text{ or } [[\text{Document}]] \text{ specifies } [[\text{Anything}]] \)

*Implicature*: \([[\text{Human}]] \text{ or } [[\text{Institution}]] \text{ or } [[\text{Document}]] \text{ or } [[\text{Concept}]] \text{ states clearly and precisely that } [[\text{Anything}]] \text{ is a necessary condition for some activity, } [[\text{Process}]], \text{ or } [[\text{Concept}]] \)

In PDEV the single instance: *A stretch of DNA which specifies a single type of protein* is marked as an exploitation

A once novel exploitation, which has been conventionalised with usage over time, perhaps motivated by the need to explain a novel concept in familiar terms.
Qualitative Analysis: *explain* \((N = 300\) in each sub-corpora)

Differences in methods across disciplines also explain differences in PT frequency:

**PT1:** \([\text{Human}]\) or \([\text{Institution}]\) or \([\text{Document}]\) or \([\text{Proposition}]\) or \([\text{Eventuality 1}]\) explains \([\text{Eventuality 2}]\) (in terms of \([\text{Eventuality 3}]\)) (to \([\text{Human 2}]\))

**Implicature:** \([\text{Human}]\) or \([\text{Institution}]\) formulates a \([\text{Proposition}]\) concerning the cause or effect of an \([\text{Eventuality 2}]\) - typically, using a *hypothesis* or theory that has already satisfactorily explained \([\text{Eventuality 3}]\)

\((= 57\% \text{ in history}; = 77.33\% \text{ in management}; = 83.33\% \text{ in microbiology})\)

Formal hypothesis testing is used much less frequently in history.
Qualitative Analysis: explain

Differences reporting practices also explain differences in PT frequency:

PT4: [[Human]] or [[Institution]] explains [Quote] (= 9 % in hist.; 6.33 % man.; = 0.33 % in microbiology) - no difference between history and management where it is used to report interview responses.

PT3: [[Human]] or [[Institution]] explains [That-Clause] (to [[Human 2]]) (= 12.33 % in hist.; 2.33 % man.; = 0.67 % in microbiology) – significant difference frequency between history and management suggests that a greater variety of structures are used for reporting in history.
Qualitative Analysis: *accept*  \( (N = 300 \text{ hist}; 157 \text{ man}; 65 \text{ micro}) \)

**Discipline-specific semantic type alternations**

**Predominant pattern:**

\([\text{[Human]}] \text{ or } [\text{Institution}] \text{ accepts } [\text{Proposition}] \text{ or } [\text{Concept}] \text{ or } [\text{Eventuality}]\)

\(= 71 \% \text{ in hist.}; 71.34 \% \text{ man.}; = 40 \% \text{ in microbiology}\)

**Implicature:**  \([\text{[Human]}] \text{ or } [\text{Institution}] \text{ agrees that } [\text{Proposition}] \text{ or } [\text{Concept}] \text{ or } [\text{Eventuality}] \text{ is correct and does not need to be contested}\)

**However in history:**

**Semantic type alternation:** \([\text{[Human]}] \text{ or } [\text{Institution}] \leftrightarrow [\text{Location}]\)

**Example:** Towns and territories **accepted** the reformation.
Qualitative Analysis: accept

Discipline-specific semantic prosody alternations

Predominant pattern:
[[Human]] or [[Institution]] accepts [[Proposition]] or [[Concept]] or [[Eventuality]]
(= 71 % in hist.; 71.34 % man.; = 40 % in microbiology)

Implicature: [[Human]] or [[Institution]] agrees that [[Proposition]] or [[Concept]]
[[Eventuality]] is correct and does not need to be contested

However in management:
Predominant semantic prosody: [[Eventuality = Negative]]
Example: Employees are more accepting of an unfavourable outcome

SP = “consistent aura of meaning with which a form is imbued by its collocates” (Louw, 1993, p. 157)
Qualitative Analysis: *accept*

**Discipline-specific syntactic alternations**

**Frequent pattern:**
[[Human]] or [[Institution]] accepts [that clause] (7 % in hist; 8 % in man.; 37 % in microbiology)

*Implicature:* [[Human]] or [[Institution]] agrees that [that clause] is true or correct

**However in microbiology:**

*Syntactic Alternation:* {it} is (generally}{well}{widely} accepted {that} - this is the only syntactic configuration present in microbiology

*Example:* It is generally accepted that compounds causing mutations in one type of cell should also be considered mutagenic for other cells.
Pedagogical Applications: Form

Adlers. Here I want to apply some of the insights that flow from:

* [[HUMAN]]
  * [[CONCEPT]]

It magistrates ought to apply the use of oaths reverently.

* [[HUMAN]]
  * [[PROCESS]]

1 months later and apply for a loan. Business loans also fell

* [[HUMAN]]
  * [for] [[BENEFIT]]

This was fruitfully applied by Huguenot authors like Mornay,

* [[HUMAN]]
  * [[CONCEPT]]
Pedagogical Applications: Benefits

Whatever form applications take they will promote reflection and the noticing of:

• **Alternations which predominate in certain disciplines:** Syntactic: e.g. It is {it} is (generally) well (widely) accepted [that]; Semantic: transparent e.g. [[Human]] with [[Location]] or more opaque: [[MB Entity]] with [[Human]].

• **Patterns which predominate in certain disciplines:** E.g. manage in spite of difficult circumstances (hist.); treat ≈ discuss (hist.)

• **Discipline-specific patterns:** [[MB Entity]] replicates (Self)
Apply
1. a person OR organisation applies theory OR model OR process SOMETIMES to address a situation
reason-of-state literature, may have been applied to practical politics. But the crucial [HISTORY]
This has made it difficult to effectively apply and combine theoretical lenses, to delineate [MANAGEMENT]
discouraged, bacteriologists did not attempt to apply the Biological Species Concept (or any [MICROBIOLOGY]

2. a theory OR model OR process applies SOMETIMES to a situation OR activity
“certain “natural”’ principles of morality” apply even without government or religion to [HISTORY]
diversification also does not apply to a single-business CEO. Accordingly, [MANAGEMENT]

3. a person applies for benefit OR an injunction OR an extension OR an admission OR readmission
of supporters and economic resources in applying for exemptions. The revised law would not [HISTORY]
upon further promotion. In 1809 Liverpool applied to the prince regent for a grant of £1,000 [HISTORY]
list. Firms included on the WM list must apply for recognition. This raises the possibility [MANAGEMENT]

4. a sample OR culture is applied to a surface OFTEN a test medium SOMETIMES by a scientist [MICROBIOLOGY]
The dialysed sample was applied to a hydroxyapatite column
100 μl of culture was applied to the face of the wedge array channel
700 μl of the final concentrated sample was applied to DNA extraction

5. a person applies a word OR term to anything
“Solidarity” was a term that could and did apply to different, overlapping, and sometimes [HISTORY]
Menu is used here as a metaphor that can be applied to a wide range of service interactions [MANAGEMENT]
The term Trojan horse has been applied to many biological packages, such as mobile [MICROBIOLOGY]

6. a person OR an organisation applies pressure
the crown initially applied pressure to compel religious orthodoxy [HISTORY]
viral mutagens apply less selective pressure to a viral population [MICROBIOLOGY]
efficient. For example, many companies have applied accountability pressures on managers to [MANAGEMENT]

A corpus-driven dictionary-like resource provides a solution to several of the problems faced by dictionary makers:

• It avoids the problems of defining technical vocabulary restricted defining vocabulary.
• Corpus examples bridge the gap between traditional definitions and examples and offer the possibility of further context from source documents.

Users are not limited to searching from word → meaning...
Multi-discipline Onomasiological Function

Function: **reporting what has been said or written**

**Explain**
- a person OR organisation explains that... SOMETIMES to another person

[Note for Historians]
Presenting the draft to Scott, Lindsay explained that its intention was to inspire respect [HIST] objections against the Xiongnu marriage code, he explains that their purpose is to safeguard the [HIST] experience. The cover sheet for the case explained that the task was to diagnose the causes [MANAGEMENT]

[Note for Management Students]
"domestic agriculture was of a purely subsistence nature“ he explained to a local intendant” [HIST] George Hulme explained: ’I think what started me thinking about this was when...“ [HIST] as one shop manager explained: “If you look after my client, when I'm away...“ [MANAGEMENT]

**Note**
- a person notes "something that is said or written using quotation marks" OR that... OR what happens OR is said or written, when, where, why something happens OR is written

the sister recalling this story for the chronicle noted: “no more widows in our convent” [HIST] participate in this study. It should be noted that research suggests that 71% of firms [MANAGEMENT] of 50 samples by RT-PCR . These authors noted that while no correlation could be established [MICROBIOLOGY]

**Specify**
- a person OR organisation OR [theory] OR [model] OR document specifies that...

[Not found in the microbiology sources used to make this dictionary]
labour‘. The Vagrancy Act went further, specifying that the ‘labour’ to be undertaken by slaves [HISTORY] he specified that the “cowries come from the Maldives“ [HISTORY] group from another” (Hofstede, 1984: 21). It specifies how things are to be evaluated and what [MANAGEMENT]
Discipline-specific Onomasiological Function

Function: *talking about the creation of cells or DNA*

**Replicate**
a cell OR part of a cell OR piece of DNA replicates another cell OR part of a cell OR piece of DNA

reporter (R) and quencher (Q) dyes. As DNA pol *replicates* the template strand, hydrolysis of the [MICROBIOLOGY] organisms are able to efficiently and faithfully *replicate* their DNA are of critical importance in [MICROBIOLOGY] P proteins to transcribe and eventually *replicate* the HPIV genome. The P protein of HPIV [MICROBIOLOGY]

**Specify**
a genome OR cell OR part of a cell OR piece of DNA specifies another cell OR part of a cell OR piece of DNA

the genes busAA-AB, *specifying* a high-affinity betaine uptake system [MICROBIOLOGY] arcABD1C1C2TD2 in which arcD1 and arcD2 *specify* putative arginine ornithine (1:1) antiporter [MICROBIOLOGY] Gene expression levels are usually *specified* by particular promoter sequences and their [MICROBIOLOGY]
Summary

• Verbs exhibit discipline-specific meanings in academic writing.
• Verb meaning is conditioned by phraseology.
  • This is a limitation of discrete-item general EAP wordlists.

• Many differences in meaning would not have been revealed by distributional approaches.
  • It is important to consider both syntactic and semantic concerns when selecting vocabulary.

• CPA-tagged data can be employed in pedagogical applications.
Thank you

g.rees@surrey.ac.uk  
@ReesGP

Thesis: https://www.tdx.cat/handle/10803/543839
References


