Clusters and Correspondences.
A comparison of two exploratory statistical techniques for semantic description

Dylan Glynn

University of Leuven
RU Quantitative Lexicology and Variational Linguistics
Aim of Study

Compare two simple techniques for exploratory multivariate analysis of semantic structure

Show that quantitative semantic analysis is possible
Cognitive Linguistics

Symbolic unit
Form-meaning pairs - no formal modules
(Langacker 1987, Fillmore & al. 1988)

Encyclopaedic semantics
No semantic modules – meaning is all conception and perception
(Fillmore 1985, Lakoff 1987)

Entrenchment
No grammar – language is usage
…no language system, social *langue*, or individual competence
Quantitative Approaches to Semantic Structure within Cognitive Linguistics

Polysemy
Lexical –
  Gries (2006) run
  Glynn (2008) hassle

Synonymy
Constructional –
  Gries (1999) VPCxs,
  Heylen (2005) Middle Field Cxs,
  Grondelaers & al. (2007) 'there' Cxs
  Speelman & Geeraerts (forth.) Causative Cxs

Lexical –
  Divjak (2006) intend verbs,
  Divjak & Gries (2006) try verbs,
  Newman & Rice (2004) posture verbs
  Newman & Rice (2004) prepositions
Hierarchical Cluster Analysis

HCA shows grouping

2-way tables

agglomerative

distance matrix

possibility of significance testing (via bootstrapping)

HCA visualisation

dendograms

different distance measures = emphasis different groupings

discrete groups = misleading semantic description
Cluster Analysis
Multiple Correspondence Analysis

MCA shows correlations

- $n$-way tables
- canonical correlation
- distance matrix

MCA visualisation

- correspondence maps
- proximity = correlation
- conflated multiple spaces = misleading proximity
Multiple Correspondence Analysis
Corpus and Annotation

LiveJournal Corpus

Online personal diaries
Very large, unparsed
British vs. American is distinguished, but little register variation
Some gender bias toward woman, probably restricted to middle class, 15-25 year olds.

Annotation

3 parameters- Semantic, Formal, and Social
120 values
20 variables
2000 occurrences
Annoy, Bother, Hassle
Breaking Down Lemmata

Transitive
  Saw quite a few people I knew, including the awful stalker guy who's been hassling me ...

Transitive Oblique
  If you hassle me about my kinky hair, I'll cut it all off. hat in hand, humble, almost begging .

Intransitive
  Officer McCoy, me and him was  hassling and my gun went off, hitting him somewhere ...

Nominal Mass
  ... because it saves all that ammoying hassle of SOD'S-BLOODY-LAW!!!!!!

Nominal Count
  I rarely paint my nails(It can be such a hassle!)

Adjective Attributive
  It's a very hassily event to do.

Adjective Predicative
  She will not take part in Saturday's 5000m race, saying she is tired and bothered

Gerund
  the technical know-how to do this sort of hassling ...
### Annoy, Bother, Hassle

#### Breaking Down Lemmata

<table>
<thead>
<tr>
<th>Form</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count Noun hassle (hassle_count)</td>
<td>146</td>
</tr>
<tr>
<td>Mass Noun hassle (hassle_mass)</td>
<td>217</td>
</tr>
<tr>
<td>Gerund hassle (hassle_gerund)</td>
<td>40</td>
</tr>
<tr>
<td>Predicative Adjective bother (bother_pred)</td>
<td>124</td>
</tr>
<tr>
<td>Intransitive bother (bother_intrans)</td>
<td>222</td>
</tr>
<tr>
<td>Transitive annoy (annoy_trans)</td>
<td>449</td>
</tr>
<tr>
<td>Transitive hassle (hassle_trans)</td>
<td>274</td>
</tr>
<tr>
<td>Transitive bother (bother_trans)</td>
<td>275</td>
</tr>
</tbody>
</table>
**Annoy, Bother, Hassle**

**Indirect Semantic Variable: Agent Type**

**Agent Type**

- Human Specific
  
  so im hassling you instead of your mum, haha!

- Human Non-Specific
  
  but we started to have more people hassling us.

- Institution
  
  Well, the Church bothers me quite often,

- Activity - Event
  
  It bothers me everytime by boyfreind talks to, or about his ex girlfriends

- Thing
  
  I pulled it out but the mouse annoys me too much...

- Abstract State of Affairs
  
  I have been open to him about everything else except that part.. however, it bothers me and I'm caught in between
Annoy, Bother, Hassle

Agglomerative Hierarchical Cluster Analysis (Dist: Euclidean/ Met: Average)
Annoy, Bother, Hassle

Multiple Correspondence Analysis

Construction-Lexeme Agent Type
Annoy, Bother, Hassle
Agglomerative Hierarchical Cluster Analysis "pvclust"

2 kinds of p-values:

AU (Approximately Unbiased)

determined by multiscale bootstrap resampling

BP (Bootstrap Probability) value

determined by normal bootstrap resampling.
Annoy, Bother, Hassle

PV Agglomerative Hierarchical Cluster Analysis (Dist: Euclidean/ Met: Ward)
**Annoy, Bother, Hassle**

Direct Semantic Variables: Cause, Affect, Humour

**Cause of Event**
- expenditure of energy
- imposition
- imposition / request
- interruption
- request
- condemnation
- tease

**Affect on Patient**
- anger
- repetition / boring
- concern - thought
- emotional pain
- physical pain

**Humour**
- Use of humour in the example
- No use of humour in the example
Annoy, Bother, Hassle

Agglomerative Hierarchical Cluster Analysis (Dist: Euclidean/ Met: Average)

Construction-Lexeme Dialect Cause Affect Humour – less forms
Annoy, Bother, Hassle

Multiple Correspondence Analysis

Construction-Lexeme Dialect Cause Affect Humour - less forms
Annoy, Bother, Hassle
PV Agglomerative Hierarchical Cluster Analysis (Dist: Euclidean/ Met: Ward)

Construction-Lexeme Cause Affect - less forms
Annoy, Bother, Hassle

Bivariate Correspondence Analysis

Construction-Lexeme Cause Affect - less forms
Russian Adjectival Constructions
Discrepancies between HCA and MCA
Russian Adjectival Constructions
Discrepancies between HCA and MCA

Cluster Dendrogram

dist_clust1_mat
hclust (*, "complete")
Annoy, Bother, Hassle

Bivariate Correspondence Analysis

Construction-Lexeme Cause Affect - less forms
Annoy, Bother, Hassle
Detail of Correspondence Analysis

Usage Cluster 1
Class Form
  Transitive *annoy*
  Transitive *bother*

Affect Features
  anger
  repetition
  concern thought
  emotional pain
  physical pain
  interruption
  aesthetic

![Graph showing the relationship between annoy, bother, and related affect features.](image)
**Usage Cluster 2**

Class Forms
- Transitive *hassle*

Cause - Affect Features
- imposition
- request
- imposition request
- tease
- condemn
**Annoy, Bother, Hassle**

Detail of Correspondence Analysis

**Usage Cluster 3**

Class Forms
- Count Noun *hassle*
- Mass Noun *hassle*
- Gerund *hassle*
- Adjective *bother*
- Intransitive *bother*

Affect Features
- energy
- agitation
Summary

Pros and Cons for HCA and MCA in Quantitative Approaches to Cog. Sem.

**HCA - groups usage patterns relative to features**

+ Possibility for significance testing

+ Clear visualisations

- 'Blind' Clustering

- Discrete Grouping

**MCA - maps usage patterns relative to visualised features**

+ ‘Analogue’ representation of associations

+ Correlations visible

- Misleading visualisations

- No significance testing
Summary

Quantitative Semantic Study

A combination of formal, indirect semantic and direct semantic tagging is possible and can produce coherent verifiable results.

Although semantic analysis is more subjective than formal analysis, if we are to describe all of language, then we should also include semantic features.
for further information:
http://wwwling.arts.kuleuven.ac.be/qlvl/
http://perswww.kuleuven.be/dylan_glynn