

# Constructing a Phenomenal Corpus

Ned Letcher



THE UNIVERSITY OF  
MELBOURNE

Supervisors:

Tim Baldwin @ The University of Melbourne  
Emily Bender @ University of Washington

# Outline

**Goal:** Construct a corpus of real world language usage annotated with occurrences of linguistic phenomena.

**Linguistic phenomena:** Things that descriptive linguistics concerns itself with and are amenable to formal analysis.

- ▶ Especially syntactic and morphosyntactic phenomena
- ▶ More complex and implementationally “interesting”

**Why:** Drive techniques for automatically detecting phenomena within grammars.

HMGE paper:

[http://www.univ-orleans.fr/lifo/evenements/HMGE13/proceedings\\_HMGE13.pdf](http://www.univ-orleans.fr/lifo/evenements/HMGE13/proceedings_HMGE13.pdf)

## Why phenomena detection?

Precision grammars have poor discoverability of linguistic phenomena

Nebulous implementations

- ▶ multiple types pertain to one phenomenon
- ▶ one type constrains multiple phenomena
- ▶ difficult to document

*One does not simply grep for linguistic phenomena...*

- ▶ phenomena names abbreviated in source code
  - ▶ eg. `v_cp_le`, `subj_prd_verb`, `conj_red_cs1_m-int_rule`
- ▶ variation in terminology
  - ▶ eg. subordinate clause, dependant clause, embedded clause
- ▶ plausible analyses vary across:
  - ▶ formalisms, grammatical frameworks, grammar engineering styles

# Phenomena detection

Difficult to quickly ascertain

- ▶ if a grammar covers a particular phenomenon
- ▶ which parts of the grammar constrain the phenomenon

Applications

1. Bootstrapping grammar documentation
2. Augmenting descriptive grammars with treebanks
3. Phenomenon-based grammar navigation
  - ▶ facilitate cross-linguistic hypothesis testing
  - ▶ leveraging implemented solutions within existing grammars

# Proposed Approach

1. Parse corpus items
2. Use parser output to associate grammar components with phenomena
3. Extract phenomenon “signatures”

## Phenomenon signatures

- ▶ clusters of types that constrain the phenomenon
- ▶ or maybe clusters of TDL constraints

Predicative adjective =

{aj\_pp\_i\_er\_le, prd\_aux\_verb\_ssr, trans\_adj\_pred\_synsem}

Passive =

{v\_pas\_odlr, norm\_passive\_verb\_lr, passive\_unerg\_synsem\_min, passive\_synsem, be\_c\_was\_le}

## Desiderata for corpus

- ▶ Grammar engineering framework independent
- ▶ Exhaustively annotated
- ▶ Token-level annotations

1. " ' Tell me , Helen , ' said she , ' have you ever heard anyone whistle in the dead of the night ? '
2. A lady dressed in black and heavily veiled , who had been sitting in the window , rose as we entered .
3. " I think that I mentioned to you that the doctor kept a cheetah and a baboon .
-

# Constructing a Proof-of-concept corpus

## Aim

- ▶ Create and refine methodology for phenomenon corpus
- ▶ Produce a proof-of-concept packaged product

Corpus: 200 lines from *Sherlock Homes and the Speckled Band*

## Methodology

1. Development of annotator guidelines
2. Annotation of text
  - ▶ One annotator full 200 lines
  - ▶ Second annotator two 50 line subsets
3. Evaluation
  - ▶ Eyeballing — 1<sup>st</sup> 50 line subset
  - ▶ Inter-annotator agreement — 2<sup>nd</sup> 50 line subset
4. Refinement of guidelines
5. Packaging of corpus

# Phenomena and Annotator Guidelines

## Phenomena selected

- ▶ Passive clauses
- ▶ Interrogative clauses
- ▶ Complement clauses
- ▶ Imperative clauses
- ▶ Relative clauses

## Annotator guidelines

- ▶ Consultation of typological literature
- ▶ Development of criteria for inclusion
  - ▶ Establish the range of each phenomenon
  - ▶ Balance between cross-linguistic coverage and non-exhaustive analysis
  - ▶ eg passives: impersonal passives, indirect passives, anticausatives

Available online: <http://repository.unimelb.edu.au/10187/17611>



# Annotation

1 On glancing over my notes of the seventy odd cases in which I have during the last eight years studied the methods of my friend Sherlock Holmes, I find many tragic, some comic, a large number merely strange, but none commonplace;

2 for, working as he did rather for the love of his art than for the acquirement of wealth, he refused to associate himself with any investigation which did not tend towards the unusual, and even the fantastic.

3 Of all these varied cases, however, I can not recall any which presented more singular features than that which was associated with the well-known Surrey family of the Royleotts of Stoke Moran.

4 The events in question occurred in the early days of my association with Holmes, when we were sharing rooms as bachelors in Baker Street.

5 It is possible that I might have placed them upon record before, but a promise of secrecy was made at the time, from which I have only been freed during the last month by the untimely death of the lady to whom the pledge was given.

6 It is perhaps as well that the facts should now come to light, for I have reasons to know that there are widespread rumours as to the death of Dr. Grimesby Roylott which tend to make the matter even more terrible.

7 It was early in April in the year '83 that I woke one morning to find Sherlock Holmes standing, fully dressed, by the side of my bed.

8 He was a late riser, as a rule, and as the clock on the mantelpiece showed me that it was only a quarter-past seven, I believed myself to be dreaming, for I was myself regular in my habits.

9 "Very sorry to knock you up, Watson," said he, "but it's the common lot this morning.

10 Mrs. Hudson has been knocked up, she retorted upon me, and I on you."

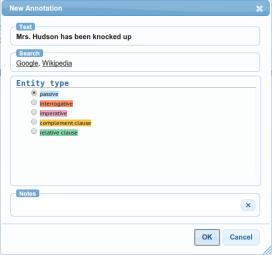
11 "What is it, then - a fire?"

12 "No;

13 a client.

14 It seems that a young lady has arrived in a considerable state of excitement, who insists upon seeing me.

15 She is waiting now in the sitting-room.



brat browser-based rapid annotation tool

# Inter-annotator agreement

## *kappa*-like coefficients

- ▶ eg: Fleiss' *kappa* and Krippendorf's *alpha*
- ▶ statistical measures of inter-rater agreement
- ▶ take into account agreement occurring by chance
- ▶ calculate agreement for phenomena spans across entire corpus

## Problems for phenomenon annotations:

1. Annotation units cannot overlap — not true of phenomena
2. Annotators are both creating units *and* labelling them
  - ▶ Introduces issues of how spans are coded
3. Want fuzzy agreement at boundaries

# Resolving overlapping units

and you know **complement clause** how subtle are the links **relative clause** which bind two souls **relative clause** which are so closely allied .

## 1. Calculate agreement on a per phenomenon basis

and you know how subtle are the links **relative clause** which bind two souls **relative clause** which are so closely allied .

## 2. Resolve nested phenomena

- ▶ For each rater:
  - ▶ For each overlapping annotation:
    1. Append copy of sentence to end of text.
    2. Move overlapping annotation to copy.
    3. Move closest matching annotation for each other rater (if any)

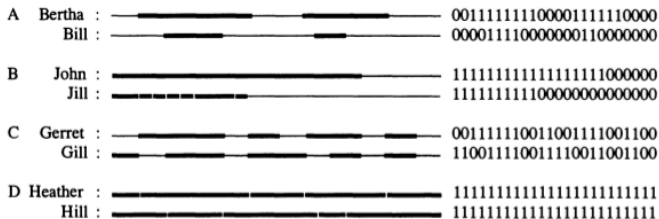
and you know how subtle are the links **relative clause** which bind two souls which are so closely allied .

and you know how subtle are the links which bind two souls **relative clause** which are so closely allied .

# Problems introduced by spanning annotations

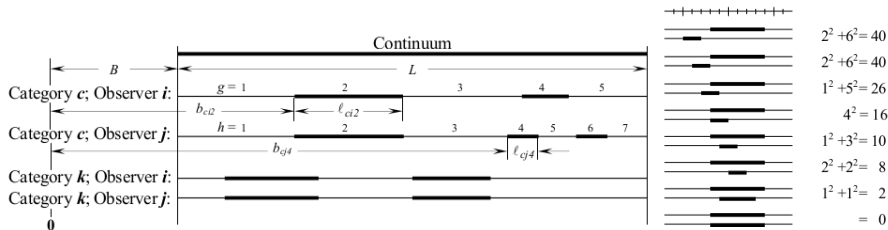
- ▶ Raters are creating units and labelling them
- ▶ Using standard approaches requires coding into tokens
- ▶ Does not respect annotation boundaries

## The coding problem



# Krippendorff's alpha for unitizing continuous data

- ▶ Divide up continuum into units: annotations and gaps
- ▶ Use a difference function between units
- ▶ Observed disagreement: compare each rater's section with all other raters' sections
- ▶ Expected disagreement: compare each possible pair of units
- ▶  $\alpha = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}}$



# Results

Phenomenon	Rater 1	Rater 2	$\alpha$ -char	$\alpha$ -word	$\alpha$ -line	$\alpha$ U-char	$\alpha$ U-word
Passive clause	4	5	0.871	0.852	0.780	0.855	0.825
Relative clause	8	8	0.909	0.910	0.854	0.888	0.899
Complement clause	8	13	0.716	0.709	0.705	0.389	0.364
Interrogative clause	3	3	0.972	0.939	1.000	0.997	0.988
Imperative clause	3	2	0.852	0.784	0.792	0.907	0.866

Interpreting kappa-like scores:

- $k = 1$  perfect agreement
- $k \geq 0.8$  reliable agreement
- $k \geq 0.667$  tentative conclusions
- $k = 0$  systematic disagreement

More work on disagreement analysis is required.

# Packaging

Export to [incr tsdb()] profile format

- ▶ Supports phenomenon and item-phenomenon records
- ▶ Existing profiles easily augmented
- ▶ But does not support character spans

## Next steps

1. Start development of phenomenon corpus based on DeepBank
  - ▶ Gold trees should yield more appropriate signatures
  - ▶ WSJ data increases interoperability between frameworks
  - ▶ ParDeepBank leaves door open for cross-linguistic exploration
  
2. Explore signature extraction techniques
  - ▶ using AVM types + supertypes from gold tree
  - ▶ Possibly complete parse forest — no need for treebank
  - ▶ Or even parse chart — no need for successful parse
  
3. Try to automate annotation using phenomenon signatures

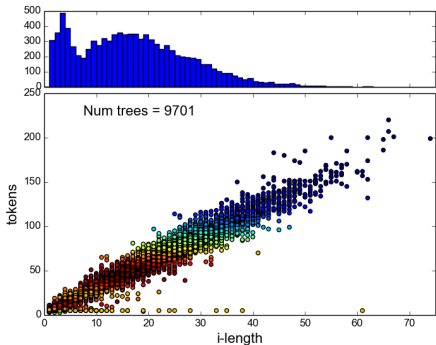


BONUS GRAPHS!

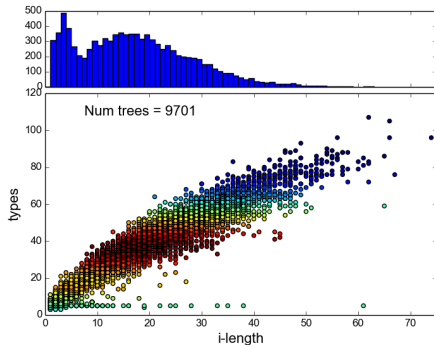
# Derivation tree tokens/types from gold trees

- ▶ rules + lex types extracted from ERG 1212 WeScience gold trees
- ▶ i-length compared to number of tokens and number of types

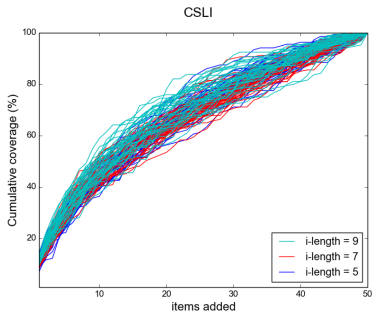
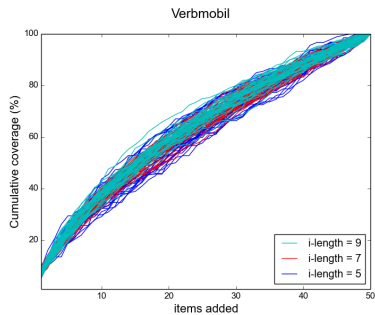
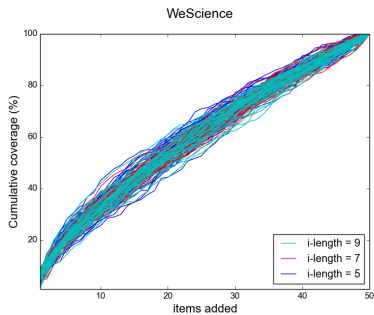
Tokens in gold trees



Types in gold trees



# Coverage of rules + lex types



# Coverage of rules + lex types

