# HG351 Corpus Linquistics

# Lexical and Grammatical Studies, Variation

Francis Bond
**Division of Linguistics and Multilingual Studies**
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 7
http://compling.hss.ntu.edu.sg/courses/hg3051/

# Overview

- Revision of DIY Corpora
  - DIY Corpora
  - Corpus Tools
  - Processing Raw Text
  - SQL
- Lexical Studies
- Grammatical Studies
- Variation

# Revision of DIY Corpora

# DIY Corpora

1. Decide what you want to study
   see if you can do it with existing resources
   No ☹

2. Collect data that fits your needs

   - Speech (expensive)
   - Text (easy to get, hard to do legally)

3. Process it

   - Clean up
   - Mark up
   - Annotation

# Web as Corpus

# Two Approaches to using the Web as a Corpus

- Direct Query: Search Engine as Query tool and WWW as corpus? (Objection: Results are not reliable)
  - Population and exact hit counts are unknown → no statistics possible.
  - Indexing does not allow to draw conclusions on the data.
  - ⊗ Google is missing functionalities that linguists / lexicographers would like to have.

- Web Sample: Use search engine to download data from the net and build a corpus from it.
  - known size and exact hit counts → statistics possible.
  - people can draw conclusions over the included text types.
  - (limited) control over the content.
  - ⊗ sparser data

# Structured Query Language

A special-purpose programming language designed for managing data held in a relational database management system (RDBMS)

- The simplest query
  - **SELECT** desired attributes
  - **FROM** one or more tables
  - **WHERE** condition applies about the records in the tables
- What lemmas are there associated with the word *does*?

| **QUERY** | **word** | **lemma** |
|---|---|---|
| SELECT word, lemma | does | do |
| FROM word | does | do |
| WHERE word = ’does’ | does | doe |
| | . . . | |

# Queries can be rather complicated

- Tell me more about frequent common nouns

```
SELECT count(word), count(DISTINCT word),
       MIN(LENGTH(word)),
       MAX(LENGTH(word)), AVG(LENGTH(word))
FROM word
WHERE word
IN (SELECT word
    FROM word
    WHERE POS in ('NN', 'NNS')
    GROUP BY word
    ORDER BY COUNT(word) DESC
    LIMIT 10)
```

| count | distinct | min | max | avg |
|---|---|---|---|---|
| 1096 | 10 | 3 | 10 | 5.13 |

- First find the ten most common words, then do things to them
  - the trick is to write simple queries first, and then combine them
- Note that more common words are shorter (as we would expect)

SQL allows you to search for a very wide variety of things

# Corpus Studies of Lexicography

# *big*, *large*, *great*

- Same syntax (all adjectives)
- Similar meaning:
  - ***large, big*** — “above average in size or number or quantity or magnitude or extent” *a large city; set out for the big city; a large sum; a big (or large) barn; a large family; big businesses; a big expenditure; a large number of newspapers; a big group of scientists; large areas of the world*
  - ***great*** — “relatively large in size or number or extent; larger than others of its kind” *a great juicy steak; a great multitude; the great auk; a great old oak; a great ocean liner; a great delay*
- How do they differ?

# Distribution of *big*, *large*, *great*

| | Academic | Fiction | Combined |
|---|---|---|---|
| ***big*** | 31 | 408 | 230 |
| ***large*** | 605 | 232 | 408 |
| ***great*** | 284 | 490 | 393 |

(frequency/million words)

- Counts from Longman-Lancaster Corpus
  - Academic Text: 2.7 Million Words
  - Fiction: 3.0 Million Words

(After table 2.5 (Biber et al., 1998, p 60))

# Immediate Right Collocates

## Academic

| *big* | | *large* | | *great* | |
|---|---|---|---|---|---|
| enough | 2.2 | number | 48.3 | deal | 44.6 |
| traders | 1.1 | numbers | 31.3 | importance | 12.5 |
| | | scale | 18.0 | number | 8.9 |
| | | and | 28.0 | majority | 8.1 |
| | | enough | 15.9 | variety | 7.0 |

## Fiction

| *big* | | *large* | | *great* | |
|---|---|---|---|---|---|
| man | 9.6 | and | 15.2 | deal | 40.4 |
| enough | 8.9 | black | 4.3 | man | 6.6 |
| and | 8.3 | enough | 3.6 | burrow | 5.6 |
| house | 7.6 | room | 2.7 | big | 4.6 |
| big | 7.0 | white | 2.7 | aunt | 4.3 |

(frequency/million words)

# Discussion *big*, *large*, *great*

- ***big*** mainly for concrete things
- ***large*** mainly for amounts and numbers
- ***great*** similar to ***large*** but many special senses
  - ***great deal***
  - ***great man***
  - ***great burrow***
  - ***great*** **`relative`**

  also use as intensifier *great big, great importance*

The dictionary definition does not really tell us this.

# Corpus Studies of Morphology

# Distribution and Function of Nominalizations

- Investigate how common normalizations are in different registers
- Count four common derivational suffixes: ***-[ts]ion, -ment, -ness, -ity***
- In three registers: Academic, Fiction, Speech
- Search for words ending in *tion, sion, ity, . . .*
  with a stop list: ***nation, station, city, . . .***
- First run the matcher, then add stop words, then rerun, . . .
  hard to do with a web interface

# Results (1)

**Nominalizations per thousand words across registers**

| Academic | Fiction | Speech |
|---|---|---|
| 44.0 | 11.2 | 11.3 |

- Nominalizations much more common in Academic text
- A few words very common (more than 500 per million)
  *movement, activity, information,* development, relation, equation*
  - If ***movement*** has occurred recently . . . (Academic)
  - Garth [. . . ] ***moved*** his hand crabwise along the table. (Fiction)
  - When we ***moved*** into the new house . . . (Speech)
- Academic text focuses on generalized processes
- Speech and fiction focus on a specific person doing some activity

(After table 3.1 (Biber et al., 1998, p 60))

# Results (2)

**Proportions of different suffixes across registers**

| suffix | Academic | Fiction | Speech |
|---|---|---|---|
| ***-[ts]ion*** | 68% | 51% | 56% |
| ***-ment*** | 15% | 21% | 24% |
| ***-ness*** | 2% | 13% | 5% |
| ***-ity*** | 15% | 15% | 15% |

- ***-[ts]ion*** more common in Academic (but common everywhere)
- ***-ment*** commoner in Fiction and Speech
- ***-ness*** common in Fiction

(After table 3.2 (Biber et al., 1998, p 63))

# Discussion

- ***-[ts]ion*** more common in Academic (but common everywhere)
  basic use is to make an action non-agentive
  - *It provides a direct indication of fuel consumption.*

- ***-ment*** often used for mental states
  *agreement, amazement, embarrassment* (Fiction)
  - *Patrick shrugged in embarrassment.*

- ***-ness*** used for personal qualities
  *bitterness, happiness, politeness* (Fiction)
  - *The bitterness in his heart was mixed with . . . .*

It would be good if we could automatically divide the words according to their semantic field (which we can approximate with WordNet, . . . )

# Corpus Studies of Syntax

# *begin* vs *start*

- ***begin*** and ***start*** are very similar in meaning
  - ***get down, begin, get, start out, start, set about, set out, commence*** — "take the first step or steps in carrying out an action" *We began working at dawn; Who will start?; Get working as soon as the sun rises!; The first tourists began to arrive in Cambodia; He began early in the day; Let's get down to work now*
  - ***begin, start*** — "have a beginning, in a temporal, spatial, or evaluative sense" *The DMZ begins right over the hill; The second movement begins after the Allegro; Prices for these homes start at $250,000*
  - ***begin, lead off, start, commence*** — "set in motion, cause to start" *the U.S. started a war in the Middle East; the Iraqis began hostilities; begin a new chapter in your life*

# *begin* vs *start*

- ***begin*** and ***start*** are very similar in possible usage
  - **intransitive**
    *I had better rest before we begin/start*
  - **transitive (NP)**
    *I will begin/start the lecture at 18:00*
  - **transitive (VP:ing)**
    *I will begin/start lecturing at 18:00*
  - **transitive (VP:to)**
    *I will begin/start to lecture at 18:00*

- So how do they differ?

# *begin* vs *start*

- Automatically tag text from two registers (Longman-Lancaster Fiction and Academic)
  `V (ADV)? NP` ⇒ T (transitive)
  `V (ADV)? to` ⇒ TCLS (to clause transitive)
  `V (ADV)? V+ing` ⇒ ING (-ing clause transitive)
  **else**: ⇒ I (intransitive)

- Aim for 250 samples, take every third

- Hand correct the initial sample

# Example

```
00018.FCT
<valency=TCLS (I)
hath her in thrall. "After a minute, the trio
==> began
rather carefully to cross the room


00021.FCT
<valency=ING (I)
station, shops, roadhouses, all closed. A dog
==> began
barking and , having begun , went on.
```

# Corrected Results

| | Intransitive | +NP | +to | +ing |
|---|---|---|---|---|
| ***begin*** | | | | |
| Fiction | 22% | 3% | 72% | 4% |
| Academic | 43% | 12% | 34% | 12% |
| ***start*** | | | | |
| Fiction | 40% | 22% | 20% | 18% |
| Academic | 64% | 16% | 15% | 6% |

- ***start*** is more common as intransitive
- ***begin*** is more common as ***to***-transitive

(After table 4.3 (Biber et al., 1998, p 98))

# Discussion

Typically ***start*** is used to show the onset of a process, often with an adverb

- *The soil formation process may start again in the fresh material*
- *The train started down the hill*

***begin*** is used with more concrete agents

- *Then I began to laugh a bit.*
- *The original mass of gas cooled and began to contract.*

Because the corpus doesn't mark **animacy** or **concrete agent** these statements are weak: we can't really make predictions or measure correlation.

# Can we do better?

- Treebanks exist for some languages
- We can search some English treebanks

```
//VP/VB/begin[->S/VP/TO/to]
//VP/VB/start[->S/VP/VBG]
```

- This can also be done offline to get counts

# What about SQL?

We can look at a word and the next word by **joining** a table to itself

- Transitive (V N)

```
SELECT a.word, b.word, b.pos
FROM word AS a JOIN word AS b
  ON a.sid=b.sid AND a.wid=b.wid-1
WHERE a.lemma='start' AND b.pos GLOB 'N*'
```

- Transitive (VP:ing)

```
SELECT a.word, b.word, b.pos
FROM word AS a JOIN word AS b
  ON a.sid=b.sid AND a.wid=b.wid-1
WHERE a.lemma='start' AND b.pos='VBG'
```

- Transitive (VP:to)

```
SELECT a.word, b.word, b.pos
FROM word AS a JOIN word AS b
  ON a.sid=b.sid AND a.wid=b.wid-1
WHERE a.lemma='start' AND b.pos='TO'
```

- Intransitive (remainder)

Regular expressions are better for this, SQL is not very good at one or none. But it is easy to write a few queries and add them together.

- Transitive (V ADV N) (none in eng.db)

```
SELECT a.word, b.word, b.pos
FROM word AS a JOIN word AS b JOIN word as c
  ON a.sid=b.sid AND a.wid=b.wid-1 AND a.wid=c.wid-2
WHERE a.lemma='start' AND b.pos GLOB 'R*' AND c.pos GLOB 'N*
```

# JOINS

An SQL **JOIN** clause is used to combine rows from two or more tables, based on common fields between them.

- (INNER) JOIN: Returns all rows with a match in BOTH tables
- LEFT JOIN: Return all rows from the left table, and matched rows from the right table
- RIGHT JOIN: Return all rows from the right table, and matched rows from the left table
- FULL JOIN: Return all rows with a match in EITHER table

```
SELECT column_name(s)
FROM table1
JOIN table2
ON table1.column_name=table2.column_name;
```

# *little* vs *small*

- ***little*** and ***small*** are nearly synonymous
- WordNet 3.0 has them share 4 synsets out of 10 for ***small*** and 8 for ***little***
  - ***small, little*** — “limited or below average in number or quantity or magnitude or extent” *a little dining room; a little house; a small car; a little (or small) group*
  - ***little, small*** — “(of children and animals) young, immature” *what a big little boy you are; small children*
  - ***little, minuscule, small*** — “lowercase” *little a; small a; e.e.cummings’s poetry is written all in minuscule letters*
  - ***little, small*** — “(of a voice) faint” *a little voice; a still small voice*
- Yet they differ semantically and syntactically

# Syntax: predicative vs attributive

- **Predicative**
  *When I was little/small, I couldn't say "hospital"*

- **Attributive**
  *It's only a little/small puppy*

- Are they used in the same way?
  - 5 million words of conversation from BNC
  - 5 million words of academic text from Longman-Lancaster

# How to find usage examples?

- Automatic pass (collect data matching patterns)
- Hand checking of a sample
- Re-weigh counts

# Automatic pass

- Match patterns against the corpus
  - **Predicative**
    *When I was {little/small}, I couldn't say "hospital"*
    `be (ADV)? (little|small)`
  - **Attributive**
    *It's only a little/small puppy*
    `(little|small) (ADJ)? NN`
  - **No tag** (the remainder)

- Store the results

```
Type = Atrb; File = 00116.TEC
section at the center of each lesion is a
-----> small
bronchus containing lungworms and ...
```

# Initial Results

| Type | Word | Atrb | Pred | No Tag | Total |
|---|---|---|---|---|---|
| Conversation | little | 2,101 | 104 | 405 | 2,610 |
| | small | 399 | 72 | 158 | 629 |
| Academic | little | 1,033 | 65 | 411 | 1,509 |
| | small | 2,557 | 316 | 399 | 3,272 |
| Total | | 6,090 | 557 | 1,373 | 8,020 |

(After table 4.1 (Biber et al., 1998, p 91))

- More **no tag** than **predicative**
- So we can't be confident
- Look at a sample (about a hundred) of each group

# Some example errors

```
Type = No tag; File = 00116.TEC  -> attributive
Shut up you
-----> little
... cow

Type = No tag; File = 00117.TEC  -> predicative
It is by no means
-----> small
for a brachiapod

Type = No tag; File = 02316.TEC  -> attributive
A: Cause they have
-----> little
B: We
A: milk bottles
```

# Hand checking of a sample

1. Extract a random sample of occurrences
   the bigger the better, make sure it is uniform

2. Analyze the grammatical feature by hand

3. Compute the proportional use of each variant in the sample

4. Multiply the total number of occurrences by these proportions

5. Adjust the original counts by the weighted counts

# Hand checking of a sample

- In this case look at a hundred from each group (6 samples)
  - Consider ***little*** in conversation
    - attributive: 100% atrb
    - predicative: 42% atrb; 39% pred; 19% other
    - no tag: 57% atrb; 4% pred; 39% other

- Use these proportions to recalculate:
  - Attributive: $2,101 + .42 \times 104 + .57 \times 405 = 2,376$
  - Predicative: $.39 \times 104 + .04 \times 405 = 57$

# Re-weighted Counts

| Type | Word | Original % Pred | Weighted % Pred |
|---|---|---|---|
| Conversation | little | 5 | 2 |
| | small | 15 | 23 |
| Academic | little | 6 | $< 1$ |
| | small | 11 | 13 |

(After table 4.2 (Biber et al., 1998, p 93))

- Adjusted counts more accurate
- Only had to check 600 (of 8,020)
- But hard to reuse or go further: only a small accurate sample parsed text is better

# Interpretation

- Attributive much more common for both
  - Predicative relatively more common in conversation
  - Predicative relatively more common for ***small*** than ***little***
- Collocation results:
  - ***little***: concrete objects (*little boy*)
  - ***small***: amounts (*small proportion*)
- But predicative ***small*** also for physical size:
  - *She's small and really skinny*
  - *He's really small isn't he?*
- We still don't really know why ☹
  corpus linguistics gives us the what, but not the why

# Can we do better?

- Treebanks exist for some languages
- We can search some English treebanks (wikipedia)
  - `//VP/ADJP/JJ/small` (predicative)
  - `//NP/ADJP/JJ/small` + `//NP/JJ/small` (attributive)
- This can also be done offline to get counts

# Where do we go from here?

- Corpora show clearly that even semantically very similar words can show different behavior.
- But they still don’t explain why
  - Hand correction limits data sizes
  - Without semantic tags, we can’t generalize automatically
- Corpora with more mark-up (syntax and semantics) would help
  - But they are expensive, . . .

# Acknowledgments

➢ Many examples from chapters 3 and 4 of Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. CUP