The dog.
The excited dog.

The man.
The man walks.

The man walks the excited dog.
As the man walks the excited dog, he daydreams of the coming spring, and is filled with dread, as he is every year when the days drag on longer, the happy sun grinning sardonically at him as he enters his windowless workplace prison for the most hectic and stressful time of year.

Example based on lecture notes of Marti Hearst, 2006

Visualizing Language

CPSC 599.28/601.28
INFORMATION VISUALIZATION

25 March, 2008

Christopher Collins
ccollins@cs.utoronto.ca
Learning Objectives

By the end of this lesson you will be able to:

- List two challenges and two advantages for text data
- Describe three key thematic areas of linguistic visualization

‘Language’

- **Text**
  - Content, form, temporality, repetition, semantics, origin

- **Speech**
  - Content, tone, melody, temporality, semantics, origin

- **Other forms of language**:
  - Sign languages
  - Sets of commonly known symbols (semiotics)
  - Mixed data (e.g., text & social networks, music)
Why Visualize Language?

- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data

Themescape
Visual Thesaurus
Thread Arcs
Visualizing Language is Difficult

- Many of the common challenges still exist
- Can you name some?

- Screen real estate / occlusion
- Choosing appropriate visual variable mappings
- Colour and perception issues
- Maintaining “graphical integrity”
- Interaction and usability

- Specific challenges for language?
Difficult Data

- Too much data – what to use?
  - Millions of blog posts,
  - Hundreds of thousands of news stories,
  - 183 billion emails,
  - ... per day
- Data is noisy:
  - Newswire stories are syndicated (but differ slightly)
  - 70-72% of email is spam
  - Text contains section headings, figure captions, and direct quotes

Once you have the data...

- Most meaning comes from our minds and common understanding.

- “How much is that doggy in the window?”
  - how much: social system of barter and trade (not the size of the dog)
  - “doggy” implies childlike, plaintive, probably cannot do the purchasing on their own
  - “in the window” implies behind a store window, not really inside a window, requires notion of window shopping

  (Hearst, 2006)
Language is Ambiguous

- Words and phrases can have many meanings, determined by context and world knowledge.
- Interesting language is often figurative:
  - “Tables encourage casual interaction.”
  vs
  - “I encouraged her to take a day off.”

I saw Pathfinder on Mars with a telescope.

Pathfinder photographed Mars.

The Pathfinder photographed Mars our perception of a lifeless planet.

The Pathfinder photograph from Ford has arrived.

The Pathfinder forded the river without marring its paint job.

(Hearst, 2006)
Data Processing Decisions

- Many levels of data processing can take place:
  - Word counting
  - Stemming: “reads” and “reading” -> “read”
  - Parsing: “USA invaded Iraq” -> Invaded(USA,IRAQ)
  - Summarization
  - Sentiment analysis: “Electronics shops have terrible customer service” -> negative assessment
  - Artificial Intelligence: “We go to the bank to obtain a loan to purchase a boat” -> bank=financial institution(72%)

- Each step of extra processing introduces uncertainty and takes time = trade-off

Data Processing is Task Dependent

- What are the information needs?

  El español es la lengua más hablada del mundo tras el chino mandarín por el número de hablantes que los tienen como lengua materna.

  Spanish is the most spoken language in the world following Chinese Mandarin by the number of speakers that have it as their mother tongue.

  ...
Tasks

1. Find lowest confidence Spanish words across models
2. Find best overall translation by probability
3. Find best overall translation by quality
4. Discover areas of disagreement between models
5. Investigate differences in order between languages

In pairs, take 3 minutes and discuss which tasks your visualization would or would not support.
Remarks on Critiques

- Concerns regarding scalability
- Category colour scales for ordered colours
- Loss of original sentences (context)
- Words far apart and disconnected make for poor readability of translations

Note: probabilities tell which system is more confident not better.
Supporters of Martin, who has been jailed without trial for more than two years, are calling on Prime Minister Stephen Harper to ask Mexican president Felipe Calderon to release Martin under a section of the Mexican constitution that allows the government to expel undesirables from the country. Martin's supporters believe she has no chance of a fair trial in Mexico. Neither does Waage.
Visual Considerations

- Text readability is dependent on size, orientation, font, clutter...
- More likely to need large amounts of text in language visualization
Visualizing language is also easy!

- SO much data available for analysis
- (Mostly) readily computer readable
- Simple techniques can give instant summaries

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>god</td>
<td>3370</td>
<td>lord</td>
<td>7872</td>
<td></td>
</tr>
<tr>
<td>we</td>
<td>1971</td>
<td>god</td>
<td>4690</td>
<td></td>
</tr>
<tr>
<td>you</td>
<td>1544</td>
<td>me</td>
<td>4096</td>
<td></td>
</tr>
<tr>
<td>lord</td>
<td>955</td>
<td>son</td>
<td>3464</td>
<td></td>
</tr>
<tr>
<td>hath</td>
<td>951</td>
<td>do</td>
<td>2979</td>
<td></td>
</tr>
<tr>
<td>your</td>
<td>845</td>
<td>you</td>
<td>2614</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>826</td>
<td>man</td>
<td>2613</td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>670</td>
<td>king</td>
<td>2600</td>
<td></td>
</tr>
<tr>
<td>make</td>
<td>549</td>
<td>israel</td>
<td>2565</td>
<td></td>
</tr>
<tr>
<td>day</td>
<td>539</td>
<td>make</td>
<td>2491</td>
<td></td>
</tr>
<tr>
<td>sura</td>
<td>538</td>
<td>hath</td>
<td>2264</td>
<td></td>
</tr>
<tr>
<td>give</td>
<td>502</td>
<td>house</td>
<td>2160</td>
<td></td>
</tr>
<tr>
<td>us</td>
<td>481</td>
<td>people</td>
<td>2139</td>
<td></td>
</tr>
<tr>
<td>verily</td>
<td>479</td>
<td>child</td>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>me</td>
<td>460</td>
<td>give</td>
<td>1872</td>
<td></td>
</tr>
<tr>
<td>send</td>
<td>437</td>
<td>we</td>
<td>1844</td>
<td></td>
</tr>
<tr>
<td>people</td>
<td>432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>earth</td>
<td>420</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Information Retrieval
Information Retrieval

- Visual query formation
- Exploration of collections
- Single/comparative document content visualization
Visual Query Formation

- Rich specification of linguistic constraints

VQuery

(Jones et al., 1998)
Exploration of Collections

- Provide overview of:
  - entire collection
  - subset matching a query
- Clustering and categorization

Galaxies

(Wise, 1999)
Tile Bars

Term Set 1: law legal attorney lawsuit
Term Set 2: network lan

(Hearst, 1995)

Lighthouse Cross-Lingual Search

(Leuski et al., 2003)
Does it work?

- NIRVE study: 3D, 2D, & text clusters
- Results depended on task:
  - Find specific title: text better
  - Find biggest cluster: 2D/text equal
  - 3D always worst!
- No one has beaten Google’s mini-summaries for standard search task

Document Content Visualization

- What are the main themes?
- How does one document differ from another?
- Where are the parts that interest me?

Consider:
What level of data processing was done?
What tasks will each visualization support?

(Sebrechts et al., 1999)
Document Lens

(Robertson and Mackinlay, 1993)

TextArc

- [http://www.textarc.org/Hamlet2.html](http://www.textarc.org/Hamlet2.html)

(Paley, 2002)
Document Icons

- Word counts
- Automatic groupings
- Drill-down

(Frid-Jimenez et al., 2005)
“Total Interaction” Essays

(Rembold & Spath, 2004)

DocuBurst

- Word counts
- Human-created ontology
- Drill-down

(Collins, 2006)
DocuBurst

- games → game
- taken → take
- absolute, noun, 10
- chair, noun, 2
- moment, noun, 11
- game, noun, 30
- reality, noun, 3
- take, verb, 13
- represent, verb, 17

... game IS activity
chair IS furniture

(Collins, 2006)
Many Eyes Tag Clouds

(Wattenberg et. al, 2006)

Linguistic Analysis
Linguistic Analysis

- Discover interesting things about languages and language use in:
  - Linguistic research
  - Language interfaces
  - Literary analysis
  - Plagiarism detection
  - Authorship attribution

Linguistic Research: Algorithm Development

Translation Parse Trees (Vockler, 2008)
Linguistic Research: Spelling Variation

German Historical Spelling Variants

(Kemken et al., 2007)

Linguistic Research: Language Structure

VisLink

(Collins and Carpendale, 2007)
New Language Interfaces

Lattice Uncertainty Visualization

(Collins et al., 2007)

Literary Analysis: Semantics

http://noraproject.org/nora_ol_video/

Nora Viz
Literary Analysis: Rhyme & Rhythm

Byron, 2007

Literary Analysis: Patterns

Don et al., 2007
Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)

Literary Analysis: Patterns

NY Times

State of the Union (Werschkul, 2007)
Senator Clinton, on the issue of jobs, I watched you the other day with your economic blueprint in Wisconsin saying, the economy is stimulating and sending a message to the state. And I've had the chance to read the monthly closely. Either say that you pledged to create million new jobs over 10 years. And I was reminded of your campaign in 2000 in Buffalo, my hometown, just three hours down Route 9, where you pledged 100,000 new jobs to update New York. There's been a notice of 30,000 jobs. And when people are asked about your pledge, you're committed. Would The Gotham\textregistered\textregistered News, "Even though we've been a little cautious..." Tonight, I think you say that the pledge of 1 million jobs might be a little optimistic. (Script Analyzer, 2007)

Clinton: No, Tim, because what happened in 2000 is that in a short All Gore was going to be president. And when I made the pledge, I was counting on having a Democratic White House. In a Democratic presidency, you shared my values about what we need to do to make the economy work for everyone and create jobs. And I've said, you know, despite the challenges of the Bush administration and the Republican Congress for six years of my 10 years. It has not been easy to create jobs, but obviously, as president, I'll have a lot more tools at my disposal. And the reason why we can create 1 million new jobs--I mean, that's not a leap. Twenty-and-one million new jobs were created during the eight years of the

(Russert, 2008)
Literary Analysis: Repetition

love is a
drug
stranger

Many Eyes Word Tree
(Wattenberg et al., 2007)
Plagiarism Detection

Two occurrences of a sequence is suspect

Authorship Attribution

Hamilton  Madison  Disputed

Federalist Papers  (Kjell et al., 1992; 1994)
Mixed Data

Language in Mixed Data Visualization

- Language as a communication medium in social networks
  - Email, IM, blogs
- Language as a label
  - Photos, videos
- Language as evidence
  - Data forensics, intelligence analysis
Social Patterns from Text

(Harris, 2006)

(Viégas et al., 2004)
Language as a Label

(Wattenberg, 2005)
Name a challenge for linguistic visualization that you encountered in last week’s exercise.

In which of the three main areas of linguistic visualization did last week’s exercise fall?
Summary

- Linguistic visualization for:
  - information retrieval
  - linguistic analysis
  - mixed-data visualization
- Challenges include:
  - amount of data
  - ambiguity of language
  - data processing decisions and constraints
  - designing for the task
  - perception of text

Action Plan

You now have the tools to incorporate some simple language visualization into your assignment three.

Contact me with feedback or questions:

cmcslin@cpsc.ucalgary.ca / www.christophercollins.ca