The *richer* the people, the *bigger* the crates
the more *literal* the *better*
the less *elaborate* you can be the *better*

The *slower* the *better*

the *higher* the temperature the *darker* the malt

The *older* a database is, the *richer* it becomes

The *higher* the clay content, the more *sticky* the soil

The *fresher* the soot, the *longer* it needed

the *higher* the number, the *greater* the protection

The *older* you are the *greater* the chance

The *smaller* the mesh the more *expensive* the net

Linguistic Institute 2017

Corpus Linguistics

Introduction

Amir Zeldes

amir.zeldes@georgetown.edu

Nathan Schneider

Nathan.schneider@georgetown.edu
Other relevant courses here in Lexington

**Historical Sociolinguistics Toolkit**

- **Dialectology** (Joe Salmons): Monday/Thursday 9:00-10:50 AM
- **Historical Sociolinguistics** (Mark Richard Lauersdorf): Tuesday/Friday 3:30-5:20 PM
- **Introduction to Historical Linguistics** (Brian Joseph): Monday/Thursday 3:30-5:20 PM
- **Introduction to Sociolinguistics** (Penelope Eckert): Tuesday/Friday 1:30-3:20 PM
- **Perceptual Dialectology** (Dennis Preston): Tuesday/Friday 9:00-10:50 AM
Other relevant courses here in Lexington

Computational / Corpus Courses

- **Combinatory Categorial Grammar: An Introduction** *(Mark Steedman)*, Tuesday/Friday 3:30-5:20 PM
- **Computational Phonology** *(Giorgio Magri)*, Monday/Thursday 11:00 AM-12:50 PM
- **Intonation and Computation** *(Julia Hirschberg)*, Monday/Thursday 3:30-5:20 PM
- **Introduction to Computational Linguistics** *(Sandra Kuebler)*, Tuesday/Friday 1:30-3:20 PM
- **Psycholinguistic and Corpus Approaches to Code-Switching** *(Melinda Fricke)*, Monday/Thursday 3:30-5:20 PM
- **Lexicography** *(Helene Schmolz)*, Tuesday/Friday 9:00-10:50 AM
Assignments

- Reading / writing assignments
- Annotation assignments
  - We will create our own corpus 😊
  - Add annotations as the course progresses
  - (Results will be published online if you’re interested!)
- We will work on annotation assignments in class (usually in part 2 of each session)
Online materials

- Reading and other assignments managed over the Canvas course: https://canvas.instructure.com/courses/1148613
  - Make sure you are registered
  - All messages will be sent over Canvas
- We will also work with other software over the Corpus Linguistics server at Georgetown University
Corpus Linguistics

- Corpus linguistics is a methodology dealing with:
  - Building
  - Annotating
  - Evaluating

Linguistic corpora

- Corpora are principled collections of language data:
  - Written text, spoken language, even multimodal recordings of sign language, and analyses thereof

- Key issue: what criteria do we use to in/exclude data?
Why corpora?

- As a very brief illustration, consider the question: what is the difference between *clever* and *intelligent*?

- And *smart*?

- Some questions are much easier to approach with corpora than without.

<table>
<thead>
<tr>
<th>and/or</th>
<th>4955</th>
<th>10052</th>
<th>2.2</th>
<th>3.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>perceptive</td>
<td>0</td>
<td>34</td>
<td>0.0</td>
<td>6.4</td>
</tr>
<tr>
<td>thought-provoking</td>
<td>0</td>
<td>32</td>
<td>0.0</td>
<td>6.2</td>
</tr>
<tr>
<td>informed</td>
<td>0</td>
<td>36</td>
<td>0.0</td>
<td>6.2</td>
</tr>
<tr>
<td>autonomous</td>
<td>0</td>
<td>46</td>
<td>0.0</td>
<td>5.2</td>
</tr>
<tr>
<td>adaptive</td>
<td>0</td>
<td>39</td>
<td>0.0</td>
<td>6.1</td>
</tr>
<tr>
<td>well-informed</td>
<td>0</td>
<td>24</td>
<td>0.0</td>
<td>6.0</td>
</tr>
<tr>
<td>literate</td>
<td>0</td>
<td>26</td>
<td>0.0</td>
<td>5.9</td>
</tr>
<tr>
<td>compassionate</td>
<td>0</td>
<td>27</td>
<td>0.0</td>
<td>5.9</td>
</tr>
<tr>
<td>well-educated</td>
<td>0</td>
<td>17</td>
<td>0.0</td>
<td>5.7</td>
</tr>
<tr>
<td>cultured</td>
<td>0</td>
<td>19</td>
<td>0.0</td>
<td>5.7</td>
</tr>
<tr>
<td>rational</td>
<td>0</td>
<td>46</td>
<td>0.0</td>
<td>5.6</td>
</tr>
<tr>
<td>playful</td>
<td>0</td>
<td>22</td>
<td>0.0</td>
<td>5.6</td>
</tr>
<tr>
<td>sensitive</td>
<td>8</td>
<td>134</td>
<td>2.0</td>
<td>5.9</td>
</tr>
<tr>
<td>thoughtful</td>
<td>14</td>
<td>121</td>
<td>5.0</td>
<td>7.7</td>
</tr>
<tr>
<td>affectionate</td>
<td>6</td>
<td>31</td>
<td>4.5</td>
<td>6.2</td>
</tr>
<tr>
<td>sophisticated</td>
<td>23</td>
<td>75</td>
<td>4.2</td>
<td>5.7</td>
</tr>
<tr>
<td>charming</td>
<td>21</td>
<td>50</td>
<td>4.9</td>
<td>5.9</td>
</tr>
<tr>
<td>insightful</td>
<td>11</td>
<td>31</td>
<td>5.2</td>
<td>6.1</td>
</tr>
<tr>
<td>resourceful</td>
<td>12</td>
<td>29</td>
<td>5.8</td>
<td>6.3</td>
</tr>
<tr>
<td>witty</td>
<td>132</td>
<td>166</td>
<td>8.3</td>
<td>8.2</td>
</tr>
<tr>
<td>inventive</td>
<td>22</td>
<td>24</td>
<td>5.9</td>
<td>5.5</td>
</tr>
<tr>
<td>clever</td>
<td>54</td>
<td>30</td>
<td>5.8</td>
<td>4.8</td>
</tr>
<tr>
<td>funny</td>
<td>233</td>
<td>103</td>
<td>7.0</td>
<td>5.7</td>
</tr>
<tr>
<td>cunning</td>
<td>16</td>
<td>7</td>
<td>5.9</td>
<td>4.0</td>
</tr>
<tr>
<td>catchy</td>
<td>19</td>
<td>0</td>
<td>5.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifier</th>
<th>4950</th>
<th>3158</th>
<th>0.9</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>emotionally</td>
<td>0</td>
<td>111</td>
<td>0.0</td>
<td>8.6</td>
</tr>
<tr>
<td>artificially</td>
<td>0</td>
<td>52</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>fiercely</td>
<td>0</td>
<td>28</td>
<td>0.0</td>
<td>7.0</td>
</tr>
<tr>
<td>highly</td>
<td>0</td>
<td>57</td>
<td>0.0</td>
<td>6.9</td>
</tr>
<tr>
<td>ferociously</td>
<td>0</td>
<td>8</td>
<td>0.0</td>
<td>6.2</td>
</tr>
<tr>
<td>supposedly</td>
<td>0</td>
<td>29</td>
<td>0.0</td>
<td>6.2</td>
</tr>
<tr>
<td>averagely</td>
<td>0</td>
<td>7</td>
<td>0.0</td>
<td>6.1</td>
</tr>
<tr>
<td>moderately</td>
<td>0</td>
<td>11</td>
<td>0.0</td>
<td>5.7</td>
</tr>
<tr>
<td>reasonably</td>
<td>0</td>
<td>54</td>
<td>0.0</td>
<td>5.7</td>
</tr>
<tr>
<td>computationally</td>
<td>0</td>
<td>8</td>
<td>0.0</td>
<td>5.6</td>
</tr>
<tr>
<td>supremely</td>
<td>0</td>
<td>7</td>
<td>0.0</td>
<td>5.5</td>
</tr>
<tr>
<td>culturally</td>
<td>0</td>
<td>12</td>
<td>0.0</td>
<td>5.5</td>
</tr>
<tr>
<td>exceptionally</td>
<td>29</td>
<td>25</td>
<td>6.0</td>
<td>5.9</td>
</tr>
<tr>
<td>remarkably</td>
<td>24</td>
<td>11</td>
<td>5.7</td>
<td>4.8</td>
</tr>
<tr>
<td>amazingly</td>
<td>17</td>
<td>7</td>
<td>5.9</td>
<td>5.0</td>
</tr>
<tr>
<td>wonderfully</td>
<td>20</td>
<td>9</td>
<td>5.4</td>
<td>4.5</td>
</tr>
<tr>
<td>very</td>
<td>1707</td>
<td>596</td>
<td>5.6</td>
<td>4.0</td>
</tr>
<tr>
<td>too</td>
<td>476</td>
<td>76</td>
<td>5.4</td>
<td>2.8</td>
</tr>
<tr>
<td>damn</td>
<td>12</td>
<td>0</td>
<td>5.6</td>
<td>0.0</td>
</tr>
<tr>
<td>dead</td>
<td>16</td>
<td>0</td>
<td>5.8</td>
<td>0.0</td>
</tr>
<tr>
<td>diabolically</td>
<td>9</td>
<td>0</td>
<td>5.9</td>
<td>0.0</td>
</tr>
<tr>
<td>awfully</td>
<td>15</td>
<td>0</td>
<td>6.1</td>
<td>0.0</td>
</tr>
<tr>
<td>terribly</td>
<td>25</td>
<td>0</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>treblyishly</td>
<td>17</td>
<td>0</td>
<td>6.8</td>
<td>0.0</td>
</tr>
<tr>
<td>torrentiously</td>
<td>16</td>
<td>0</td>
<td>6.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>10048</th>
<th>16081</th>
<th>2.0</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>being</td>
<td>0</td>
<td>208</td>
<td>0.0</td>
<td>6.1</td>
</tr>
<tr>
<td>robot</td>
<td>0</td>
<td>77</td>
<td>0.0</td>
<td>6.1</td>
</tr>
<tr>
<td>agent</td>
<td>0</td>
<td>455</td>
<td>0.4</td>
<td>6.0</td>
</tr>
<tr>
<td>guess</td>
<td>0</td>
<td>35</td>
<td>0.0</td>
<td>5.5</td>
</tr>
<tr>
<td>railing</td>
<td>0</td>
<td>27</td>
<td>0.0</td>
<td>5.3</td>
</tr>
<tr>
<td>layman</td>
<td>0</td>
<td>22</td>
<td>0.0</td>
<td>5.3</td>
</tr>
<tr>
<td>conversation</td>
<td>0</td>
<td>88</td>
<td>0.0</td>
<td>5.1</td>
</tr>
<tr>
<td>creature</td>
<td>11</td>
<td>137</td>
<td>2.4</td>
<td>5.9</td>
</tr>
<tr>
<td>lyric</td>
<td>81</td>
<td>90</td>
<td>5.8</td>
<td>5.7</td>
</tr>
<tr>
<td>fellow</td>
<td>52</td>
<td>14</td>
<td>5.1</td>
<td>3.1</td>
</tr>
<tr>
<td>pass</td>
<td>67</td>
<td>9</td>
<td>5.2</td>
<td>2.2</td>
</tr>
<tr>
<td>stuff</td>
<td>146</td>
<td>8</td>
<td>5.1</td>
<td>0.4</td>
</tr>
<tr>
<td>gimmick</td>
<td>15</td>
<td>0</td>
<td>5.1</td>
<td>0.0</td>
</tr>
<tr>
<td>satire</td>
<td>10</td>
<td>0</td>
<td>5.1</td>
<td>0.0</td>
</tr>
<tr>
<td>flick</td>
<td>21</td>
<td>0</td>
<td>5.2</td>
<td>0.0</td>
</tr>
<tr>
<td>loot</td>
<td>15</td>
<td>0</td>
<td>5.3</td>
<td>0.0</td>
</tr>
<tr>
<td>pun</td>
<td>19</td>
<td>0</td>
<td>5.3</td>
<td>0.0</td>
</tr>
<tr>
<td>ruse</td>
<td>17</td>
<td>0</td>
<td>5.5</td>
<td>0.0</td>
</tr>
<tr>
<td>eh</td>
<td>24</td>
<td>0</td>
<td>5.8</td>
<td>0.0</td>
</tr>
<tr>
<td>wordplay</td>
<td>21</td>
<td>0</td>
<td>5.8</td>
<td>0.0</td>
</tr>
<tr>
<td>chap</td>
<td>47</td>
<td>0</td>
<td>5.9</td>
<td>0.0</td>
</tr>
<tr>
<td>twist</td>
<td>94</td>
<td>0</td>
<td>6.5</td>
<td>0.0</td>
</tr>
<tr>
<td>trick</td>
<td>166</td>
<td>0</td>
<td>6.7</td>
<td>0.0</td>
</tr>
<tr>
<td>clog</td>
<td>50</td>
<td>0</td>
<td>7.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Data from Sketch Engine/UKWAC
Linguistics and data

- What kinds of data do we use in linguistics?
- How do we collect them?
- What properties do different kinds of data have?
- What kind of data do we need in order to answer what kinds of questions?
Linguistics and data

- We can use a variety of data sources (also concurrently)

- Some examples:
  - Introspection
  - Questionnaires
  - Corpora
  - Psycholinguistic experiments
  - ...
Introspection

- Armchair Linguistics? (we’ll discuss Fillmore 1992 on this!)
- Generative competence model:
  - What are the possible utterances of the language $L$ (Chomsky 1957:13)
  - Use of grammaticality judgments
  - Search for a parsimonious model to derive the members of $L$

Armchair of Charles Dickens (Wikimedia)
Intuition – a quick exercise

- What determines the choice between:
  - *The speech of the president*
  - *The president's speech*

- Some suggestions:
  - Animacy of possessor: ±human, ±animate, ±abstract
  - Animacy of possessed: ±human, ±animate, ±abstract
  - (Relative?) length of possessor/possessed
  - Givenness/newness

- Which is more important than which?
Results: Gries (2002)

- **Comparison of:**
  - Professional linguists’ assessments
  - Questionnaire directed at ‘naïve’ native speakers
  - Spontaneous corpus data (spoken and written, British National Corpus)

- **Main results:**
  - All data agrees on Animate possessor → ’s genitive
  - Linguists otherwise contradictory (different judgments on relevance of length, givenness, possessed)
  - Length is empirically irrelevant (but spoken vs. written data)
  - Questionnaire largely in line with corpus
  - Corpus reveals extremely skewed frequency of inanimate possessed
Corpus-assisted armchair linguistics?

- Advantages:
  - intuition is always available
  - parameters can be varied easily (you can't always find "that same sentence but with a ...")
  - relative ratings of malformed sentences possible (Featherstone 2005)
  - correlates with corpus frequencies and more – converging evidence

- Disadvantages:
  - judgments are highly subjective, especially for the interesting cases 😊 (hence: more expensive questionnaires, experiments)
  - selection of data is subjective – harder to mitigate
  - no context or made-up context are not like naturalistic contexts
  - no quantitative data (but possibly graded judgments)
Questionnaires

- Range from "asking one's colleagues" to systematic experiments
- A good questionnaire is developed with:
  - Specific criteria: research questions, distractors ...
  - Target group(s) in mind
- Collects:
  - Grammaticality judgments
  - Elicited material (natural language?)
  - Arguably a principled collection of language data → elicited/questionnaire corpus
Psycholinguistics

- Primary question: how is language processed in the brain?
  - Reaction times
  - Capacity and accuracy of memory
  - Dealing with errors
  - Interaction with other cognitive processes...

- Methods:
  - Perceptual experiments (e.g. lexical decision task)
  - Production experiments
  - Rating experiments
  - Eye tracking, self-paced reading
  - fMRI, other imaging techniques
Data in linguistics – some polarizing stereotypes

<table>
<thead>
<tr>
<th>Grammaticality judgments</th>
<th>Psycholinguistics</th>
<th>Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence: what will be judged grammatical?</td>
<td>Processing: how does this happen?</td>
<td>Performance*: what will occur?</td>
</tr>
<tr>
<td>Generative system to produce all and only grammatical utterances</td>
<td>Model organization of language perception and production in the brain</td>
<td>Stochastic model to predict distribution of language data</td>
</tr>
<tr>
<td>Qualitative: categorical grammaticality (<em>the the</em>)</td>
<td>Qualitative: postulation and testing of modular structures</td>
<td>Qualitative: new constructions, categorically different varieties</td>
</tr>
<tr>
<td>Quantitative: graded grammaticality, experiments</td>
<td>Quantitative: arguments from reaction times, memory restrictions</td>
<td>Quantitative: probabilistic, multifactorial</td>
</tr>
</tbody>
</table>

* “Performance” can mean many things!
Competence vs Performance

- I maintain a distinction between competence and performance (some corpus linguists do not)

- For the purposes of this course:
  - Competence describes **consistent** patterns (including usage information, frequencies...)
  - Performance refers to **random** deviations from these patterns (e.g. non-systematic dysfluency)

- It turns out that very many distributional facts about language are systematic

- Big part of L1 : L2 difference (see Fillmore 1979)
Probabilistic Syntax

- As formulated by Manning (2003), one of the main goals of (quantitative) linguistics is modelling:
  - \( P(\text{meaning} | \text{utterance}, \text{context}) \)
  - \( P(\text{utterance} | \text{meaning}, \text{context}) \)
- Corresponds to Halliday’s (1977) view: Language is how to mean
- These goals cannot be approaches without corpora
Corpora

- Very heterogeneous field:
  - (Plain) text corpora (Web data, newspaper language)
  - Spoken corpora (possibly with phonetic transcription and annotations), multimodal corpora (annotation of gestures and non-verbal communication possible)...
  - Richly annotated corpora – treebanks (syntax trees for every sentence), error annotation (for L2 corpora)...
  - Multilayer corpora – combinations of multiple, independent annotation types
The right tool for the job

- Choosing/building a corpus depends on the **research question** or application in mind first.
- It is a **bad idea** to first choose a methodology and then decide what question you want to answer.
- Corpora are also not the right tool for every question:
  - They tell us what occurs and in what contexts for a certain type of language.
  - How often, and what can occur instead in those contexts.
  - How one type of language differs from another.
  - No categorical negative evidence (but some indications).
No "one size fits all"

- Corpora can be tiny (digitized piece of ancient papyrus, full of holes)
- or huge (10 billion words for some large freely available corpora of English from the Web)
- Very different types of annotation:
  - Often parts of speech
  - Sometimes syntax
  - Semantics, word senses, coreference, rhetorical structure, information structure...
Types of questions

- What kinds of questions can corpora help us to answer?
- What should the data (ideally) look like?
- How should it be analyzed?

➤ Focus on quantitative analysis – we're just counting stuff, right?
Debunking non-qualitative quantitiveness

- There is no such thing as non-qualitative quantitative work
- Quantitative does mean that we're counting something
- But you can't count cases until you've made a qualitative decision for each one
An example

- Let's count the prepositions in this text:
  - *Parents participate in the sports activities of their children more often during the summer all over America and, in fact, the world over.*

- Depending on what you're interested in, you could reach very different counts

- More important than anything:
  - Formulating explicit guidelines
  - Reproducible results (for people other that yourself!)
  - Learning how to do this is a big part of this course
Once you can classify you can count

- [http://metropho.rs/](http://metropho.rs/) - X is the Y of Z map
Some typical "quantitative" areas

Corpora figure prominently in:

- Sociolinguistics
- Dialectology
- Historical linguistics / language change
- Lexicography
- Field research
- Psycholinguistics (to evaluate / vet experimental data)
- Usage-based linguistics (as proxy for linguistic experience)
- Language acquisition (L1 & L2)
- Converging studies (with experiments, psycholinguistics)
Sociolinguistics

- What distinguishes the way different people speak
  - Phonology / phonetics
  - Lexicon
  - Syntax
- Why?
  - interpersonal variation
  - sociolect, ethnolect, gender, register...
- What do we need?
  - Data from different types of speakers
  - Metadata documenting relevant parameters
  - Statistical models of variation and change
Dialectology and language variation

- Where does this go?

B: 'A' I don't understand why IBM they advertise they use [pause] elephant uh
C: Elephant ah
B: Ya there's nothing to do with computers
C: Ya That's what ah that's how IBM think of our customers
B: What
C: Elephants will buy from us
A: Better don't say ha cannot say [pause] You had it I tell you [pause] Better censor it man
C: (makes elephant noises)
B: Terrible
**Historical linguistics**

- No native speakers for dead languages / older stages
- When did things happen and why?
  - How did people come to say *you* instead of *thou*?
  - Why didn't they do it in German (*du*)?
  - How do old words die out and new ones come about?
  - Are the universal/recurring tendencies? Can we predict change?
- What do we need?
  - Diachronic corpora, comparable texts across time (difficult!)
  - Mathematical models of language change
Language acquisition

- What are typical phases/recurring phenomena in L1/L2 acquisition? What kinds of errors do we see?
  - How do children learn to speak/write?
  - What predicts different outcomes? → pedagogy
  - Interference between multiple L1s/L2s?
  - Stylistic differences between native/foreign writing? ...

- We need:
  - Richly annotated corpora of language acquisition
  - Include comparable target data: adults/natives in maximally similar context
Experimental work / psycholinguistics

▪ How do features of words in usage affect processing?
  ▪ Does frequency matter for a certain task?
  ▪ Does frequency carry across related words (priming)?
  ▪ Can we learn something about the brain from the effects?

▪ We need:
  ▪ "Representative" corpora (how frequent is the word really in general? What does this mean?)
  ▪ Annotation of features relevant to processing (length, phonological and morphological structure...)

In rented accommodation forever. Hi Bruno. <Long time no speak>. I prefer your new identity to old ones 20 April, 2006 at 1:03 PM Hello <Long time no browse>, they have banned the "google translated" version.
Lexicography

- How is a word used?
  - Examples, relative frequency/importance of senses
  - Inclusion, exclusion and ordering of entries
  - Related words, collocations (*harbor* → *grudge*)
  - Terminology, professional language

- We need:
  - Balanced corpora (represent different varieties)
  - Syntactic annotation (crucial for distinguishing homonyms)
  - Good search tools, collocation and term extraction
Choice of corpus matters!

These people are under continual disquietudes, never enjoying a minutes peace of mind; and their disturbances proceed from causes which very little affect the rest of mortals. Their apprehensions arise from several changes they dread in the celestial bodies: for instance, that the earth, by the continual approaches of the sun towards it, must, in course of time, be absorbed, or swallowed up; that the face of the sun, will, by degrees, be encrusted with its own effluvia, and give no more light to the world; that the earth [...]  

[Gulliver’s Travels, Vol. 3 / J. Swift]
Choice of corpus matters!

Good Hot Dogs In DC, Maryland Or Virginia?

I love a good hot dog and I'm always looking for a good place to get one (when I fly out to Los Angeles, friends who pick me up at LAX know that we HAVE to go straight from the airport to Pink's so I can score a couple'a Ozzy dogs) but I have yet to find more than one or two places in this area to score a good dog.

I don't expect to find a place as good as Pink's and with such a varied menu, but I'd be happy to find someplace that knows you can do more with a dog than just put mustard, onions, chili or cheese on it.

So far I've enjoyed:

1.) Ben's Chili Bowl
2.) Fatburger (good Cali dog and onion rings)

Five Guys was okay, but even with the variety of toppings, their dogs don't impress me. Tried Ann's Dari-Creme and I was less than impressed (their rolls sucked).

So does anyone have any thoughts or suggestions?

[dslreports.com: Forums → Regional Talk → Washington & Baltimore]
Choice of corpus matters!

okay -that's fine. Now, on the investigation, you know, the Democratic break-in thing, we're back to the-in the, the problem area because the FBI is not under control, because Gray doesn't exactly know how to control them, and they have, their investigation is now leading into some productive areas, because they've been able to trace the money, not through the money itself, but through the bank, you know, sources - the banker himself. And, and it goes in some directions we don't want it to go.

Choice of corpus matters!

20 USC § 78 (2011)

§78. Cooperation of Smithsonian Institution with State institutions for continuing paleontological investigations

- The Secretary of the Smithsonian Institution is authorized to cooperate with any State, educational institution, or scientific organization in the United States for continuing paleontological investigations, and the excavation and preservation of fossil remains, in areas which will be flooded by the construction of Government dams or otherwise be made unavailable for such investigations because of such construction: Provided, That such investigations and activities shall not duplicate nor affect adversely similar operations being conducted by the Department of Interior in cooperation with the Smithsonian Institution.

[Aug. 15, 1949, ch. 427, §1, 63 Stat. 606.]
"General purpose" corpora

- Constructing corpora is expensive
- We would like to use the same resource for multiple research questions
  - Reusable resources
  - Licensing
  - Documentation
- Different design policies/philosophies
  - Reference corpora
  - Monitor corpora
  - Opportunistic corpora
Reference corpora

- Other terms: sample corpus, national corpus...
- Premise: a representative example of the language X
- Examples for English:
  - British National Corpus (BNC)
  - Corpus of Contemporary American English (COCA)
- Features:
  - Large, fixed size (occasionally newer versions), reproducible
  - Known composition (proportions: genres, periods, authors)
  - "Standard" but marginalizes non-mainstream text/speech types
  - At some point: goes out of date (e.g. Brown corpus, 1961)
Monitor corpora

- Dynamically growing corpora
- Internal composition often unstable (hard to collect new data in exactly the same proportions)
- Examples:
  - Corpus of Contemporary American English (COCA, as of 2012)
  - Bank of English (used by COBUILD dictionary)
  - Google ‘corpora’ – Google Books, N-Gram ‘corpora’
- Features:
  - Coverage of latest language
  - Good for lexicography, studying neologisms, short-term language change
  - Constant upkeep effort
Opportunistic corpora

- Get everything you can:
  - Web data, parliament proceedings
  - The Bible, Human Rights Declaration
  - Harry Potter (copyright!) or fan fiction 😊

- Examples:
  - Most Web corpora: WaCKY (Baroni et al. 2009), COW corpora (Schäfer 2015)
  - Most parallel corpora: EuroParl (Koehn 2005), Hansards (Roukos et al. 1995)

- Features:
  - Relatively cheap to collect, popular in Computational Linguistics
  - Often hard to say what they contain, imbalanced, unclear parameters (applicability of annotation schemas?)
Corpora and this course

- Which type of corpus could we build in this course?
- Some different design policies/philosophies:
  - Reference corpora
  - Monitor corpora
  - Opportunistic corpora
Building our own corpus

- Starting today we will build a real corpus of social media data over the remainder of this course:
  - The **Linguistic Institute Reddit Corpus (LIRC)**
- A richly annotated subset of the **Georgetown University Historical Reddit Corpus (GUHRC)**
- Partly modeled after the **Georgetown University Multilayer corpus (GUM)**:
  - [http://corpling.uis.georgetown.edu/gum](http://corpling.uis.georgetown.edu/gum)
Building our own corpus

- What kind of research questions can we look at?
  - Recent language change (2005-2017)
  - Trends in social media
  - Establishment of new genres
  - Subreddit communities
  - Meta data: upvotes, sticky, ...

- Applications for Computational Linguistics
  - Domain adaptation, coverage of social media
  - Train tools to address larger volume of data
Publication and licensing

- Publishing data
  - Students accomplish tremendous amounts of work in the classroom
  - More often than not ends up in a drawer/bin
  - In this course you will get a chance to **publish** your annotated corpus data (100% optional!)
Publication and licensing

- Licensing

  - Publishing is not enough – without a **license** no one is really allowed to use your data

  - Data published in this course will be under a Creative Commons Attribution license (**CC-BY**)
    - Anyone can use, modify and expand the data
    - Must give the source (you!) credit

  - We can publish data **anonymously** or under your name (metadata), or not at all!
Get your texts!

- Our texts will come from random Reddit posts that are:
  - All ~250 words long (consequences?)
  - Stratified dating between 2005-2017
  - Metadata for date, subreddit, upvotes, stickied and more...
  - Warning: Social Media is sometimes offensive... Let me know if you need a different text!

- You can find your texts here:
  - [https://corpling.uis.georgetown.edu/gitdox/institute2017/](https://corpling.uis.georgetown.edu/gitdox/institute2017/)
  - Login: first.last (lower case!)
  - Password: 12345
Finding your text in GitDox

- [https://corpling.uis.georgetown.edu/gitdox/institute2017/](https://corpling.uis.georgetown.edu/gitdox/institute2017/)

![GitDox: Project LIRC](image)
First assignment - for Tuesday

- Reading (after class):
  - Fillmore (1992), in Canvas. We will discuss in class:
    - Why do we need corpora?
    - How do they intersect with your research interests?

- First contact with our data
  - Read your text
  - Think of one word that sums it up (e.g. “horses”)
  - Add metadata tag to reflect this name: short_name
  - Split the text into sentences using <s> ... </s>
Adding metadata

- **Rules:**
  - Field name: `short_name`
  - Field value: one word
  - All lower case, only characters [a-z]

- **Mine:** `superman`
Adding sentence splitting

<text id="LIRC2017_dgys1z8"><p>Absolutely. The film Snyder doubt the original cut. The 3 hr UE.</p>
<p>I guess management needed to be a part of the production his job, but he had no oversight, but even he was aware the studio should ve just released the UE. It probably been a more respected film, and that carries more wait in its mostly contained to the last act. Having a 2 hour set up of GA. They want to be entertained. WB probably saw this and wa to do that had there been multiple action sequences through of Superman in Africa, Knightmare Batman, Batmobile scene, This film needed clear work from the very beginning. The sort of minimal blueprint for the director that sort of mat</p></text>
Adding sentence splitting

Absolutely. The film without a doubt the original cut. The 3 hr UE. I guess management needed to be a part of the production make. He did his job, but he had no oversight, but even. The studio should've just released the UE. It pre it would've been a more respected film, and that carries more action sequences. It was mostly contained to the last action much quicker, even though they wouldn't have had to do another film. But all we got in the first 2 hrs was the short scene then the non-stop 3rd hour. This film needed clear work from the very beginning. create a sort of minimal blueprint for the director that sort
What else can we add?

- We will add annotations using the GUM scheme:

- More discussion next time, but to foreshadow:
  - `<sic>` tags for errors: *carries more <sic>wait</sic> in the end*
  - Date and time annotation:
    - `<date when="2015-07-05">July 5, 2015</date>`
    - `at <time when="20:00:00">8</time>`
  - *You said <quote>“Superman sucks”</quote>*
  - Scare quotes: *So-called <q>“reasons”</q>*
References