Corpus Linguistics (L615)  
Corpus Building  

Markus Dickinson  
Department of Linguistics, Indiana University  
Fall 2015
Corpus Building

We will walk through 10 design principles for corpus building (Sinclair 2005, Tono 2003, Lozano & Mendikoetxea 2013)

▶ Then we will turn to some practical matters for corpus building (Kübler & Zinsmeister, sec. 1.2.1)

We will follow the presentation in Lozano & Mendikoetxea (2013), which itself follows Sinclair (2005).

▶ I encourage you to think critically & challenge any criteria you don’t fully agree with
What makes a corpus?

First, some general questions to consider:

▶ What is the difference between data & corpora?
▶ How do I create a corpus?
▶ How do I represent my corpus?
▶ What should be in my corpus?
▶ Do I need to annotate my corpus linguistically?
▶ How do I annotate my corpus?
▶ How do I search in a corpus?
▶ How do I distribute (/find) a corpus?
Difference between data and corpus?

Generally speaking, a corpus . . .

- is electronic
- is useful for more than one person
- has a clearly defined format
- is potentially annotated with a clearly defined annotation scheme
- has a name
 Principle 1: Content selection

Select according to external criteria, not internal

*The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function (Sinclair 2005: 1).*
Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen (Sinclair 2005: 2).
Principle 3: Contrast

Only those components of corpora which have been designed to be independently contrastive should be contrasted (Sinclair 2005: 3).

As Lozano & Mendikoetxea (2013) state, “comparisons within a corpus can be made only if the corpus has been designed to allow for such comparisons.”

- e.g., for comparisons of learner data: include equivalent native corpus, different proficiency levels, etc.
Principle 4: Structural criteria

Criteria for determining the structure of a corpus should be small in number, clearly separable from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination (Sinclair 2005: 5).

- e.g., spoken vs. written; different genres
Principle 5: Annotation

Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications (Sinclair 2005: 5).
Principle 6: Sample size

Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get to this target [as soon] as possible. This means that samples will differ substantially in size (Sinclair 2005: 7).
Principle 7: Documentation

*The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken* (Sinclair 2005: 8).
Principle 8: Balance

The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components (Sinclair 2005: 8).
Principle 9: Topic

Any control of the subject matter [i.e. topic] in a corpus should be imposed by the use of external, and not internal, criteria (Sinclair 2005: 10).
Principle 10: Homogeneity

A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided (Sinclair 2005: 14).

- Rogue texts are anomalous ones which would make the corpus less representative.
- Would defining rogue by linguistic properties get us back to internal criteria?
Practical Steps

0. Decide on what the corpus is to be about
   ▶ See also: 10 design principles

1. Obtain text (or speech) & decide on corpus structure

2. Put text into electronic form
   2.1 Correct automatic errors (OCR)
   2.2 Retain text formatting
   2.3 Ensure proper character encoding

3. Segment the text

4. Add meta-data

5. Add other annotation layers

6. Document the process


http://ota.ahds.ac.uk/documents/creating/dlc/index.htm
Obtain text

Obtaining text depends on what form it’s available in:

▶ hard copy: scan into (text-based) pdf
▶ jpg, gif, image-based pdf, or other images: convert to text-based pdf (OCR)
▶ pdf, doc, or other readable files: convert to plain text
▶ web files: strip html
  ▶ Some mark-up may be desirable to retain ...
▶ ...

The goal is to obtain **plain text** files
Decide on corpus structure

Decide on the organization of the corpus:

- One huge file?
- A large directory with a consistent file-naming scheme?
- Several sub-directories, organized by some principle (e.g., genre, author, etc.), with consistent file naming?
- ...

Other questions include:

- Separate directories for documentation & tools?
- Inclusion of README files?
- Storage & linking of meta-data?
Put text into electronic form

- Scanning (with OCR)
  - Freely available tools such as tesseract-ocr have support for different languages
    - Google “comparison shopping” could be your best bet ...
  - You will likely have manual correction to do

- Text conversion
  - Whatever character encoding the text is in, UTF-8 has become a standard for encoding
    - Whatever encoding, the corpus should be consistent
  - A tool such as iconv can perform conversion between encodings
    (http://stackoverflow.com/questions/64860/best-way-to-convert-text-files-between-character-sets)
Put text into electronic form
Retain text formatting

Major questions for text files are:

1. What text formatting do you want to retain?
2. How do you retain it?

Examples include:

- Textual markup: bolding, italics, etc.
- Non-linguistic text: formulae, tables, etc.
- Non-textual information: images, etc.

Standoff annotations & XML mark-up allow for incorporation of such elements
Segment the text

The text is more than a sequence of characters

▶ Sections, paragraphs, sentences, words (tokens), ...
▶ Adding segmentation allows for reusability among researchers

**Tokenization** is potentially the most difficult of the bunch & is related to the purpose of the text in the end:

▶ *in spite of* $\mapsto ||in\spite\of||$ or $||in||spite||of||$?
▶ *don’t* $\mapsto ||don’t||$ or $||do||n’t||$ or $||do||not||$?

Languages which have no spaces are even more challenging

**Sentence segmentation**: in English, periods are ambiguous between sentence boundaries & abbreviation markers (inter alia)
Add meta-data

Encode various bits of information about the corpus:

- Data properties (source, year, encoding, etc.)
- Authorial properties (gender, age, etc.)
- Annotator properties
- ...

How do you anticipate people using the data, & what meta-information will support that usage?

Where to encode this information depends in part upon the corpus representation (next slides)

- General corpus readme file
- Header of each corpus document
- Embedded within XML structure
Add other annotation layers

We will spend much time on this as the semester goes on ...

For now, think about not only what you want but:

- Where you will store the annotation (e.g., standoff format)
- How you will store the annotation (e.g., for complex annotations)
- How you will document the annotation
Document the process

Along the way, keep good internal documentation on decisions
  ▶ Eventually converted to externally understandable documents

You will likely want:
  ▶ Top-level README and COPYRIGHT document(s)
  ▶ A doc/ directory with guidelines, publications on the corpus, or anything else relevant

You may also want to include any software you’ve developed for processing the corpus
Example corpora

Under /Volumes/Data/en/, let’s examine various corpora to see what we can figure out, what we do/don’t like, etc.
In-class example

Let’s break into groups & design a corpus of “IU speech”
  ▶ i.e., a corpus representing how people at Indiana University (Bloomington) speak in 2015