Corpus Linguistics (L615)
Corpus Building

Markus Dickinson
Department of Linguistics, Indiana University
Spring 2013

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

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

I encourage you to think critically & challenge any criteria which you don’t fully agree with

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).

Principle 2: Representativeness

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

- e.g., for learner data, include an equivalent native corpus; include different proficiency levels to compare; 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
<table>
<thead>
<tr>
<th>Principle 5: Annotation</th>
<th>Principle 6: Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>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).</td>
<td>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).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principle 7: Documentation</th>
<th>Principle 8: Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>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).</td>
<td>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).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principle 9: Topic</th>
<th>Principle 10: Homogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>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).</td>
<td>A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided (Sinclair 2005: 14).</td>
</tr>
</tbody>
</table>
  - Rogue texts are anomalous ones which would make the corpus less representative.  
  - Would defining rogue by linguistic properties get us back to internal criteria?