Research Designs Using Content Analysis

Klaus Weber
Northwestern University
klausweber@northwestern.edu

PDW “Content Analysis in Organizational Research”

Academy of Management
Montreal, 2010
A Definition

“Content analysis is a research technique for making replicable and valid inferences from texts [broadly conceived] to the contexts of their use.”

Krippendorff, 2004

- **Technique**: Systematic (reliable, valid)
  - Evidence based, transparency of procedure

- **Text**: Data is generated in communication process
  - Author, audience, intent, context matter

- **Inference**: Interpretative in nature, inference about social reality
  - Researcher, research question, theoretical constructs

- Examples of how data is generated: Conversations, speeches, articles, open ended survey responses, interviews, images
Outline

• What’s in a Text? A Semiotic Framework
• Generic Types of Textual Analysis
• The Process From Collecting Data to Presenting Results
• Sampling, Base Rates and Control Groups
• Validity and Reliability
• Software Support
What’s in a “Text”? A Semiotic Perspective

- **Text** = words/images arranged in order, but of interest are often *ideas* or *actions* that the words point to.

- **The semiotic problem**: words, concepts, and referents do not correspond one-to-one (the good news: associations are usually conventionally defined in a particular social context).

-> Be clear about what the text analysis is getting at: Linguistic patterns, cognitive-cultural schemas, “facts”?  

**Solutions to the semiotic problem:**

- Three generic sources of meaning and interpretation:
  - **Referential**: A word’s meaning derives from its association with a referent or idea (e.g., categories, names)
  - **Relational**: Meaning derives from a signs position to other signs (e.g., association, opposition, grammar, plots)
  - **Contextual**: Meaning derives from the communication context (who created the text, for whom, when, why)

---

*Klaus Weber, Content Analysis, Academy of Management Meeting 2010*
What Content Is Analyzed

**Inductive Theory Building**
- Holistic interpretation (-> themes, mechanisms, taxonomies)

**Quantitative Content Analysis**
- Small text units in isolation, e.g. categories (-> frequencies, trends, etc.)

**Semantic Analysis**
- Relationship between content units, e.g. associations and grammar
  (-> scripts, networks of associated concepts, causal maps, narratives)

**Computational Linguistics**
- Algorithm based identification of structure, e.g. latent semantic clusters
  (-> analogue: factor and cluster analysis; problem: interpretation)

**Discourse Analysis**
- Several texts, e.g. regimes of interpretation
  (-> broad ideologies, institutional myths and political contradictions)
The General Process

<table>
<thead>
<tr>
<th>Step</th>
<th>Activities / Decision Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Generation &amp; Collection</td>
<td>Sampling, interviews, archived documents, web content, recordings of conversations &amp; speeches, etc.</td>
</tr>
<tr>
<td>Data Storage &amp; Organization</td>
<td>Digitalization (transcribing, OCR scans); Cleaning and Formatting, Unitization and Indexing (author, document, sentence, etc.)</td>
</tr>
<tr>
<td>Categorizing and Connecting</td>
<td>Category development (custom, standard); Dictionary development (custom, standard); Operational Definition of Associations</td>
</tr>
<tr>
<td>Coding and Aggregation</td>
<td>Sampling, Coding, Reliability/Validity, Aggregation of unit-level coding</td>
</tr>
<tr>
<td>Simplification and Presentation</td>
<td>Frequencies, Trends, Comparisons, Networks, Maps, MDS, Statistics, etc.</td>
</tr>
</tbody>
</table>
Sampling, Base Rates and Control Groups

- **Sampling “authors”**
  - Most common

- **Sampling texts**
  - No need to code every document (in most instances)

- **Sampling coding units**
  - No need to code every paragraph / sentence (in most instances)

- **Base rates of word or category occurrence**
  - Random is fine but often not most realistic

- **Control groups**
  - Corpus linguistics
  - Closely matched other texts
Validity and Reliability of Inference

- **Validity**: external (generalization), internal (constructs and causality)
- **Reliability**: replication over time, across individuals
  - Multiple researchers for category development
  - Applicability of generic categories and dictionaries
  - Documentation (code book, dictionary)
  - Reliability statistics: Don’t use percent agreement, avoid Cohen’s kappa; use statistics that account for chance agreement between coders: Krippendorff’s alpha, Scott’s pi
## Software Support Options: Some Packages

<table>
<thead>
<tr>
<th>Type of Software</th>
<th>Examples of popular software</th>
<th>Functionality</th>
<th>Functionality</th>
<th>Functionality</th>
<th>Functionality</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory Building Support</td>
<td>ATLAS.ti, Ethonograph, Kwalitan, MaxQDA, NVivo</td>
<td>Yes (best for smaller volumes)</td>
<td>Yes (main focus)</td>
<td>Some (best for smaller volumes)</td>
<td>Some (mostly basic)</td>
<td>Little (export to other software)</td>
</tr>
<tr>
<td>Coding Support</td>
<td>Diction, TextQuest, LIWC Yoshicoder WordStat</td>
<td>Yes</td>
<td>Little</td>
<td>Yes (main focus, efficient for high volume)</td>
<td>Little</td>
<td>Little (export of other software)</td>
</tr>
<tr>
<td>Mapping</td>
<td>AutoMap, Decision-Explorer, UCINet</td>
<td>Some</td>
<td>Little</td>
<td>Yes (main focus)</td>
<td>Yes (e.g. concept centralities)</td>
<td>Some</td>
</tr>
<tr>
<td>Text Mining</td>
<td>TextAnalyst, SAS plug-in, SPSS plug-in WordStat, TAKMI</td>
<td>Yes (especially for large volumes)</td>
<td>Little</td>
<td>Yes</td>
<td>Some (e.g. built in algorithms)</td>
<td>Some</td>
</tr>
</tbody>
</table>
A Few Key Questions To Ask

• What does the “content” represent?
• How were the data generated?
• Are categories and measurements valid and replicable?
• What is the base rate of “content” occurrence?
• What inferences / interpretations can I actually make?