
Tutorial: Computer-Supported Text Analysis

Klaus Weber
Northwestern University
klausweber@northwestern.edu

**PDW “The Power of Richness III”
Academy of Management
Philadelphia, 2007**

Software Support – A Word of Caution

Why you would want to use software support:

- Flexible: Easy to manage and re-code data, e.g., for emerging themes
- Validity: Coding rules and choices are consistent and documented
- Efficient: Enables systematic analysis of very large text corpuses
- Expedient: For observational digital data such as emails, web content

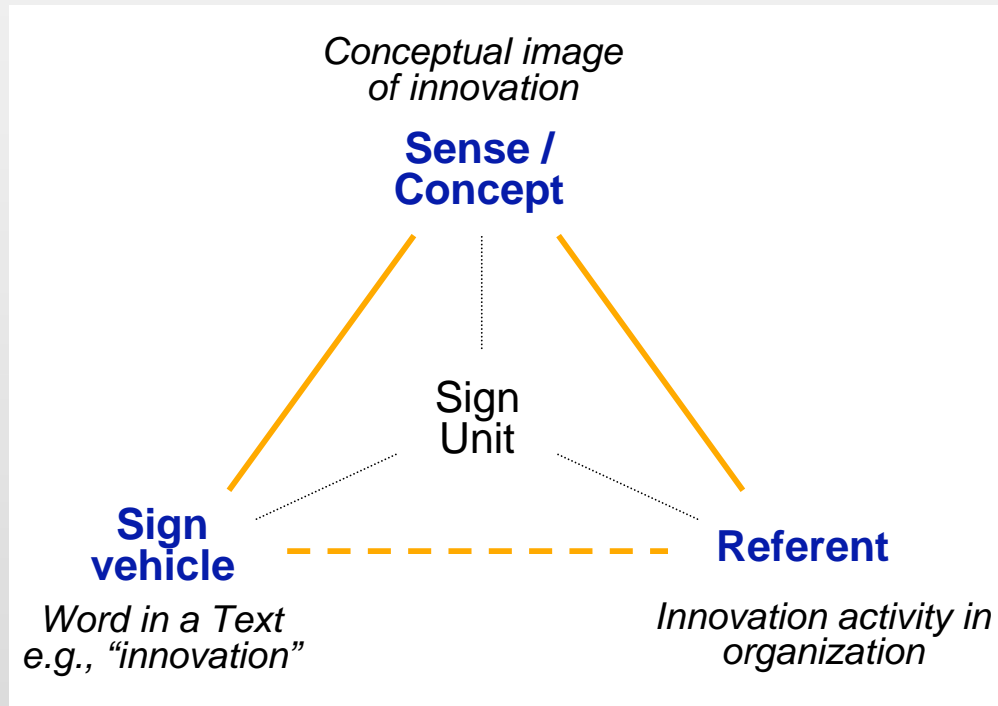
But also:

- It's no substitute for a good research design and analytic approach!
 - "Both packages offer a variety of features that effectively help researchers run associations and present results. However, in extracting themes from unstructured data, both packages were only marginally helpful. **The researcher still needs to read the data and make all the difficult decisions.**" The American Statistician 2005: 59(1): 89-103 in a comparison of *SAS Text Miner* and *WordStat* packages
 - "I used NVivo to analyze the transcripts" is about the same as saying "I used Stata to analyze the data", i.e. it says nothing about the quality of analysis
- Limited functionality, especially for complex meaning structures
- Don't underestimate the time and effort involved!
 - E.g.: OCR quality depends on the quality of the source documents (e.g., layouts, speckles); custom software needed for different steps

Outline

- **What's in a Text? A Semiotic Framework**
- **Generic Types of Textual Analysis**
- **The Process From Collecting Data to Presenting Results**
- **Illustrations From a Software-Supported Process**
- **Software Support Packages and Functionality**
- **Further Resources**

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 – and available software support:

–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 sign's 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)

Varieties of Text Analysis

Inductive Theory Building

- Holistic interpretation (-> themes, mechanisms, taxonomies)

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, etc.)

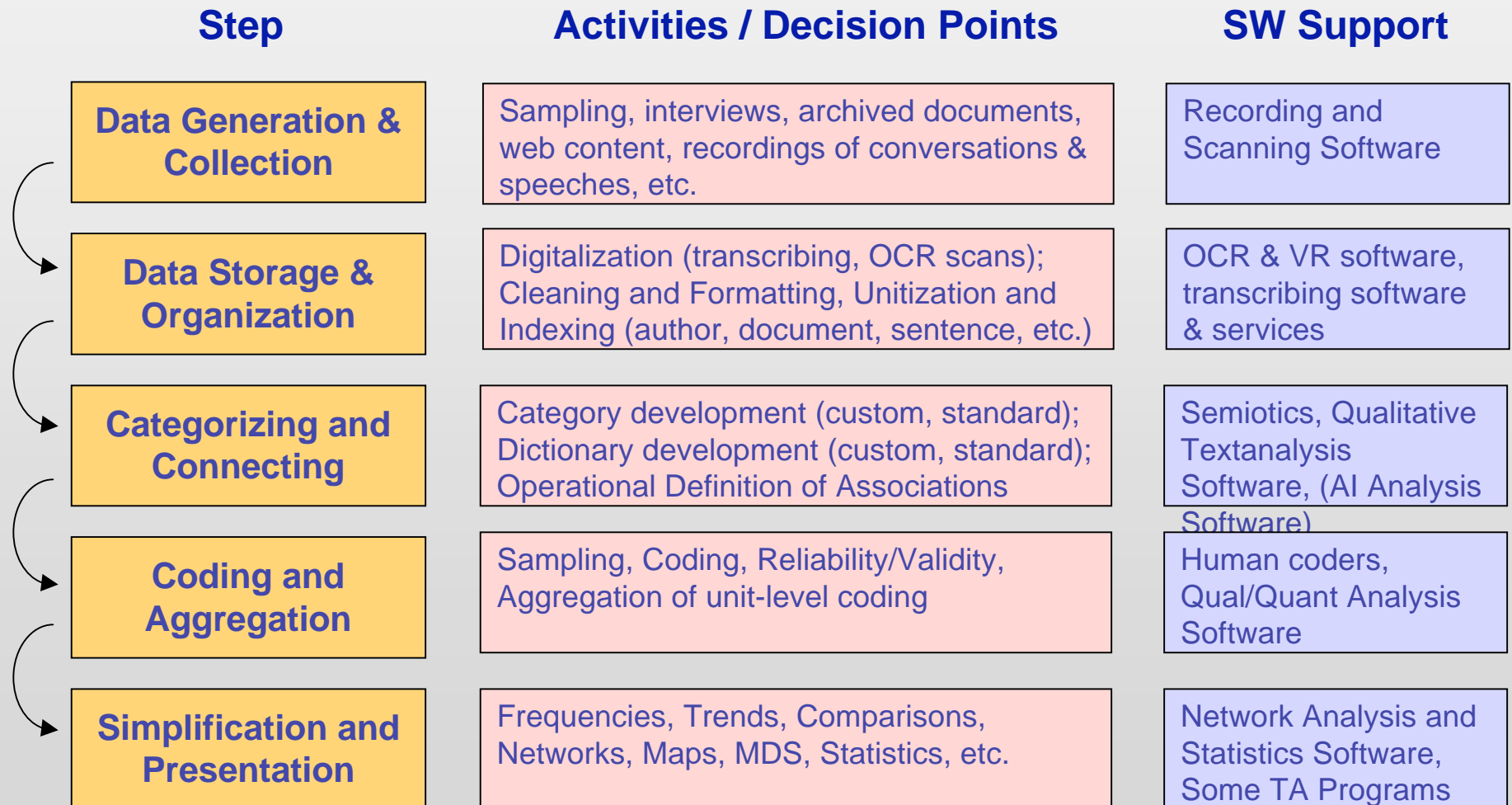
Narrative analysis

- Structure of larger text units, e.g. elements, turns, plots in a story (-> more complex stories and rhetorical practices and beliefs)

Discourse analysis

- Several texts, e.g. regimes of interpretation (-> broad ideologies, institutional myths and political contradictions)

The General Process



Dictionary
Category - Code – Phrase

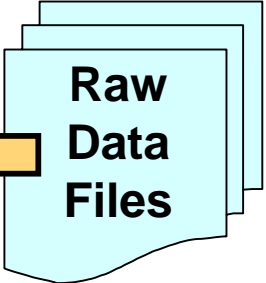
| | | |
|-----|------|-------------|
| 001 | love | “dear” |
| 001 | love | “affection” |
| 002 | hate | “hate” |

Disambiguation
Negations, exclusions

| | |
|-----|-------------------------|
| 001 | if “not” then ignore |
| 002 | if “dearborn” then drop |
| 003 | if 001 & 002 then 001 |

Text file
Organized by text units
Indexed

| | |
|----------|-------------|
| 001-B-89 | Text unit 1 |
| 002-R-89 | Text unit 2 |
| 003-A-91 | Text unit 3 |
| . | |
| . | |
| . | |
| 987-B-15 | Text unit n |



Sequential codes output

c001, c002, - , c002, c002,
....

Tabular codes output

| | |
|----------|-------------------------|
| 001-B-89 | Text unit 1: c001, c002 |
| 002-R-89 | Text unit 2: - |
| 003-A-91 | Text unit 3: c002 |
| . | |
| . | |
| . | |
| 987-B-15 | Text unit n: c002, c002 |

Illustration: Automap

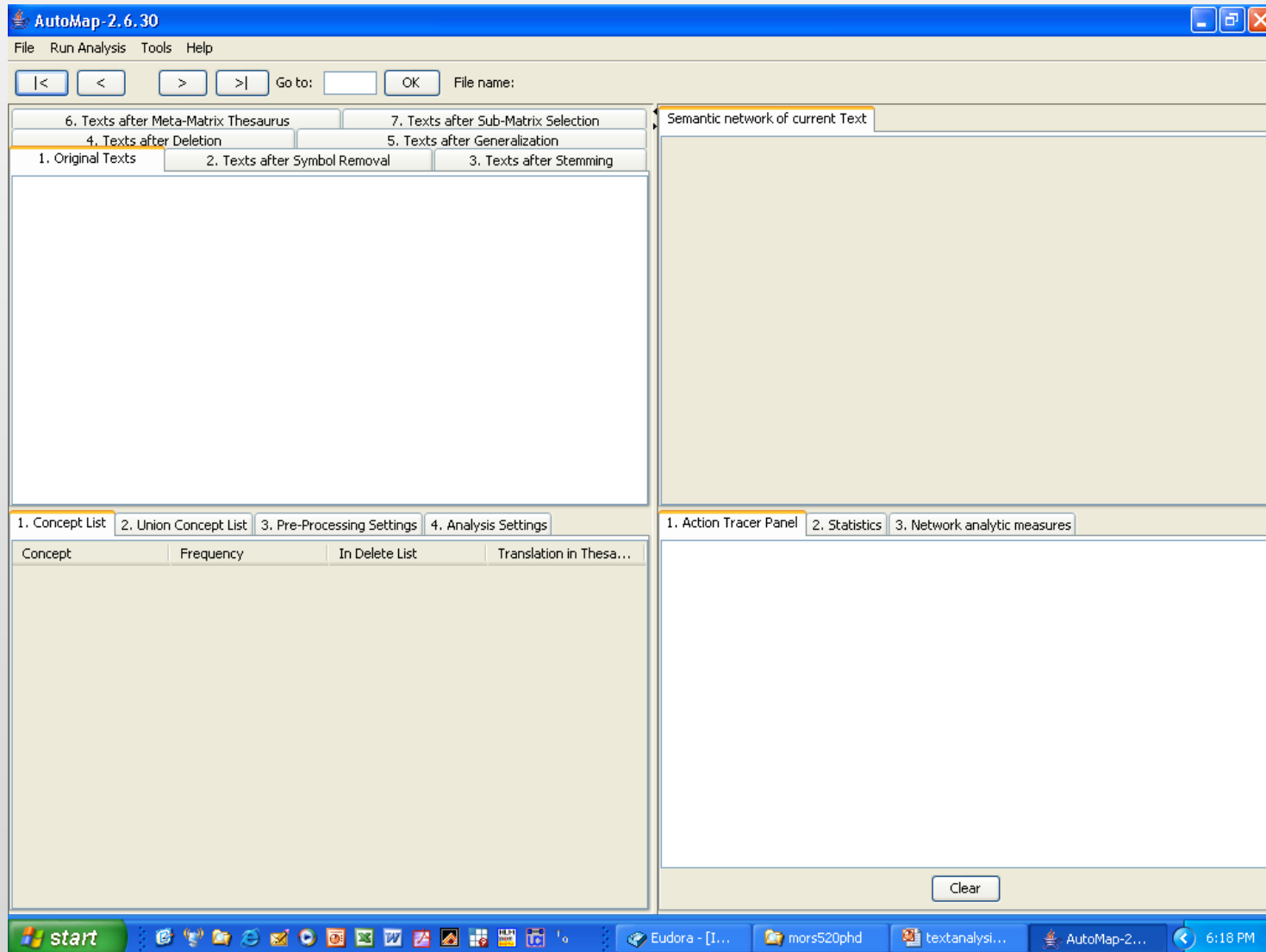
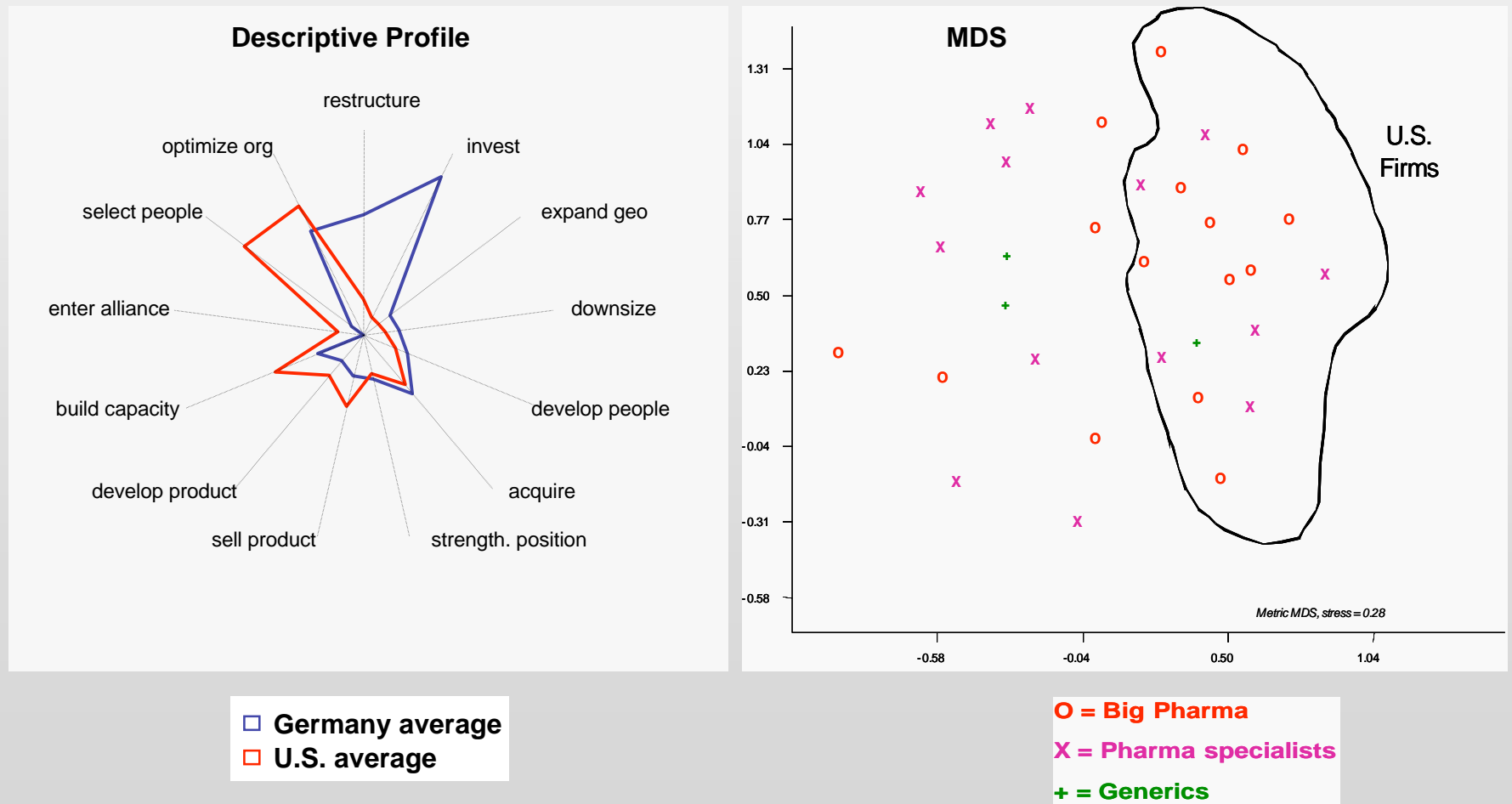


Illustration: Other Forms of Data Reduction

- Corporate Cultural Repertoires -



Software Support Options: Packages

| Type of Software | Examples of popular software | Functionality Storage, retrieval | Functionality Developing and linking categories | Functionality Automated content coding | Functionality Mapping, display of coded data | Functionality Quantification, statistics |
|-------------------------|--|--|---|---|--|---|
| Data Input | OmniPage,, FineReader, ReadIRIS, NaturallySpeak other plug-ins | Yes (single purpose) | No | No | No | No |
| Theory Building Support | ATLAS.ti, Ethonograph, Kwalitan, MaxQDA, NUD*IST/NVivo | Yes (best for smaller volumes) | Yes (main focus) | Some (best for smaller volumes) | Some (mostly basic) | Little (export to other software) |
| Coding Support | Diction, TextQuest, VbPro | Yes | Little | Yes (main focus, efficient for high volume) | Little | Little (export of other software) |
| Mapping | AutoMap, Decision-Explorer, UCINet | Some | Little | Yes | Yes (main focus) | Some (e.g. concept centralities) |
| Text Mining | TextAnalyst, SAS plug-in, SPSS plug-in WordStat, TAKMI | Yes (especially for large volumes) | Little | Yes | Some | Some (e.g. built in algorithms) |

References and Resources

A short and highly selective list of methods oriented references

- Corman, Steven R, T Kuhn, T D McPhee, and T J Dooley. 2002. "Studying complex discursive systems: centering resonance analysis of communication." *Human Communication Research* 28:157-206.
- Dohan, Daniel and M Sanchez-Jankowski. 1998. "Using computers to analyze ethnographic field data: Theoretical and practical considerations." *Annual Review of Sociology* 24:477-498.
- Eden, Colin, F Ackermann, S Cropper. 1992. The analysis of cause maps. *J. Manage. Stud.*, 29 309-324.
- Fairclough, Norman. 2003. *Analyzing discourse: textual analysis for social research*. New York: Routledge.
- Franzosi, Roberto. 1995. "Computer-assisted content analysis of newspapers: can we make an expensive research tool more efficient?" *Quality & Quantity* 29:157-172.
- Krippendorff, Klaus. 2003. *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Mohr, John W. 1998. "Measuring meaning structures." *Annual Review of Sociology* 24:345-370.
- Murmann, Johann P, Homburg, E, Geven, R, Bermiss, Y S and Forgione, A, "Automatic Coding of Printed Materials" (July 19, 2007). *SSRN working paper*. <http://ssrn.com/abstract=1001568>
- Nadkarni, S and VK Narayanan. 2005. Validity of the structural properties of text-based causal maps: An empirical investigation, *Organizational Research Methods*, 8(1): 9-40.
- Roberts, Carl W. 2000. "A conceptual framework for quantitative text analysis." *Quality & Quantity* 34:259-274.
- Weber, Klaus. 2005. "A toolkit for analyzing corporate cultural toolkits." *Poetics* 33:227-252.
- West, Mark D. 2001. "Theory, method, and practice of computer content analysis." Westport, CT: Ablex.

Some web sites with further information and links to articles and software:

http://www.slais.ubc.ca/resources/research_methods/content.htm

<http://www.car.ua.edu/>

<http://www2.chass.ncsu.edu/garson/pa765/content.htm>

http://bama.ua.edu/~wevans/content/ppp/ppp_menu.htm

<http://www.content-analysis.de/general.html>

<http://academic.csuohio.edu/kneuendorf/content/>