Matters of Fact, Opinion, and Credibility: Distinguishing Stochastic from Substantive Information in Texts

by

Carl W. Roberts

Conference on Optimal Coding of Open-Ended Survey Data

December 4-5, 2008
(University of Michigan, Ann Arbor)

Overview

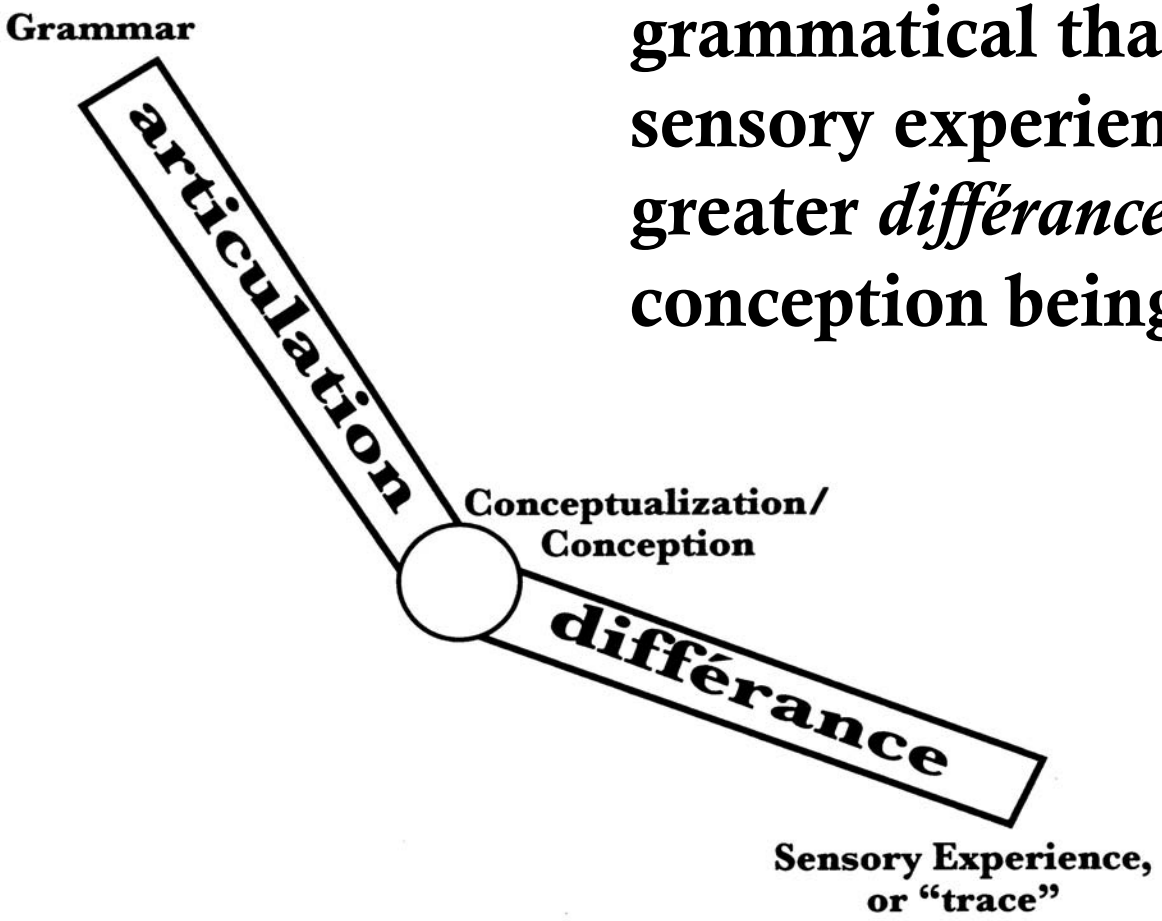
- **A nonstarter: Encoding "the" content in texts**
- **Methods for encoding texts correspond to the questions being asked.**
  - **Is the investigator's expertise assumed, or is the respondent's expertise sought?**
  - **Are inferences to be made about theme prevalence or theme relations?**
- **A (maybe fanciful) recommendation**

Assumptions

- **Language is NOT a neutral medium through which "content" is transmitted .**
- **Linguistic expressions are simultaneously…**
  - **manifestations of perspectives, and**
  - **propositions regarding sensory experience.**

Derrida's Hinge

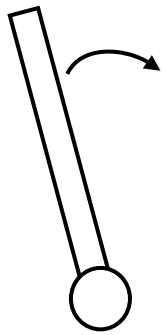
Some articulations will be more grammatical than others; some sensory experiences will have greater *différance* (sic) with the conception being conveyed.



Whose Grammar; whose Sensory Experience?

- **If the investigator (*I*) is expert …**

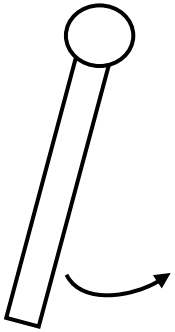
"Chief Justice"



***I*'s Sensory Experience**

- **– *I*'s expectation of "Chief Justice of the US Supreme Court" may not jibe with the sensory data, "Chief Justice," from the texts.**

***I*'s Grammar**

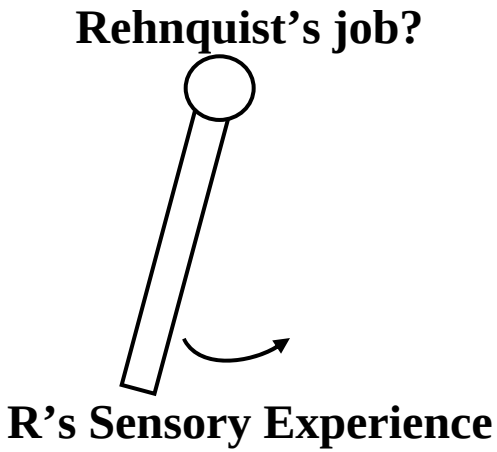


"Chief Justice"

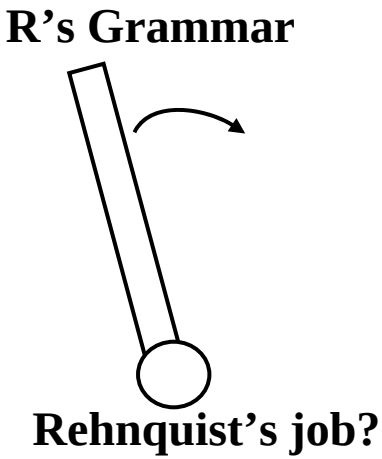
- **– *I*'s expectation of a grammatically correct reference to a senior job within the US government may be met with the same phrase.**

Whose Grammar; whose Sensory Experience? (contd.)

- If the respondent (*R*) is expert …

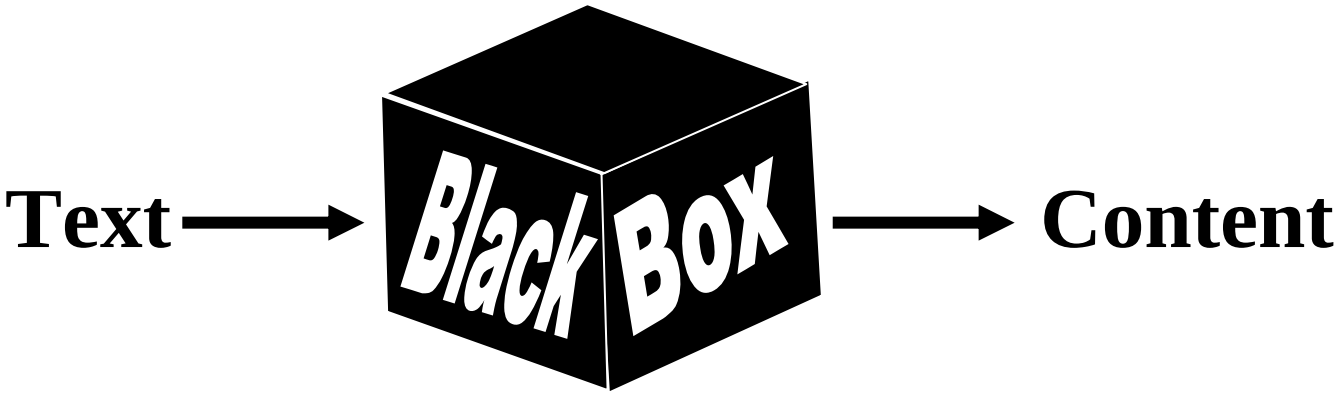


  - *R*'s experiences (as intern?) regarding Rehnquist's job might include, "I've seen him send appeals back to lower courts if the SC's justices don't agree on them."



  - *R*'s personal perspective (or grammar) regarding Rehnquist's job might be rendered as, "A job done worse than done by any of his predecessors."

Revealing questions, not content



Although a universal "black box" for revealing "the content" in texts may sound appealing, a more fruitful revelation might be of the *questions we are bringing to the texts*.

Question #1

As investigators, are we …

- **experts, interested in identifying the words respondents' use or in diagnosing how they use them, or are we …**
- **novices, interested in gaining factual or attitudinal information from respondents?**

That is, do we presume the respondent or ourselves to have the appropriate perspective for interpreting the texts at hand?

Why not simply encode the texts according to a "general audience" perspective?

- "When our Prime Minister negotiates, his first offer is his last offer!"
  - A hard bargainer?
  - Said in 2001 by (later PM) Sharon regarding Israeli Prime Minister Barak.
- "Don't blame me. I didn't vote for him…, I think." (Florida bumper sticker in early 2001)
  - Floridians are stupid?
  - Florida's officials are incompetent?
- In sum, audiences are multiperspectived; when made procedurally explicit, the investigator's or respondent's perspectives are not.

The investigator as expert

- As expert, one can automate one's expertise via fixed dictionaries, disambiguation routines, machine learning, etc.
  - The General Inquirer's dictionaries are tailored to classify texts into Osgood's semantic dimensions.
  - Machine learning algorithms classify texts in accordance with various definitions of "similarity."
- More sophisticated text analysis software incorporates parsing algorithms that encode how respondents relate the words and phrases listed in one's dictionary.
  - Louis Gottschalk's software encodes respondents' psychological states (anxiety, hostility, etc.).
  - Phil Schrodt's TABARI software was developed to parse short (e.g., Reuters) news clips.

The respondent as expert

"What do you *(as expert)* think has been the most important issue facing the United States over the last four years?"

- **Key-Word-In-Context (KWIC) software allows users to classify texts in accordance with respondents' actual vocabulary.**
- **Kristin Behfar's concept mapping methods enlist respondents themselves in the development of thematic categories.**
- **TCA and PC-ACE are for software assisted encoding of theme-relations from the respondent's perspective.**

Question #2

Are inferences to be made about theme prevalence or theme relations?

That is, are we interested in the conditions under which …

- **some themes are more prevalent than others, or**
- **themes are related in specific ways?**

(Note that the software and methodologies listed on the previous 2 slides are broken down into these two research objectives.)

Which themes are prevalent

- **Every content analysis involves—as Nigel Fielding correctly points out—some data reduction, if only by collapsing responses into thematic categories.**
- **Based on word/phrase counts within these categories, researchers may legitimately draw inferences about the prevalence of themes among various types of texts.**

Which themes are prevalent (contd.)

- Such data are often subjected to a **co-occurrence analysis** (i.e., a calculation of correlations among word-counts), and, unfortunately, to inferences about how themes are related in text windows (e.g., of 10 consecutive words).
- In making these inferences researchers commit the **ecological fallacy** of assuming clause-level relations based on text-window-level associations.
- The bottom line: If one's research question deals with grammatical relations among words, these relations must be encoded directly at the outset.

How themes are related

- National cross-sectional surveys might be used to reconstruct facts that respondents may have observed (e.g., who criticizes whom within the family unit).
- Roberto Franzosi's actor-action-object semantic grammar might be useful in capturing such relational information.

How themes are related (contd.)

- More commonly, national cross-sectional surveys are used to access respondents' perspectives (e.g., as reflected in their "most important issue facing the US").
- My own work on semantic grammars for investigating cultural perspectives has potential relevance here.
- A brief illustration of semantically encoded data might be instructive at this point.

Encoding relational information on respondent perspectives

- Consider the 1960s follow-up question to "the most important issue facing the US":

  The question: "What would you like to see done about this problem?"

- Note that responses to this question call for relational encoding.
  - A possible response: "We need to keep jobs in the United States."
  - Its semantic elements: ["We"(subject)+"keep"(verb)+ "US jobs"(object)] is necessary(modality).

A data matrix of encoded theme-relations

ID	Subject	Verb	Object	Modality
1*	11	35	64	1
2	9	22	89	3
3	14	35	72	2
4	11	36	72	1
5	17	30	55	4
				

* Rendering: We (11) keep (35) US jobs (64) is necessary (1)

Methodological notes

- **The ecological fallacy is avoided when such semantically-encoded data are used in making inferences about how themes are related in texts.**
- **Semantically-encoded data may not be simultaneously used to analyze both the facts (as experienced by the respondent) and the respondent's perspective. Let me explain.**

Countervailing biases in respondents' expertise

- Note that in conveying the facts they have witnessed, differences among the perspectives respondents use in interpreting these facts (e.g., animosity toward some family members) may bias their renderings of these facts.
- Moreover, if respondents' words are considered expressions of their perspectives, differences in the facts of their lives (job loss, divorce, etc.) may bias their rendering of these perspectives.
- Thus whereas respondents' perspectives introduce noise into their reports of the facts, such facts introduce noise into renderings of their perspectives.

Summary

- My focus has been on two key questions that shape the types of content encoded from texts:
  - Is the investigator's or the respondent's expertise being presumed?
  - Are inferences to be made about theme prevalence or theme relations?
- Answers to these questions impose substantive structure on one's data—structure independent of stochastic variation among responses.
- Since these questions have no "correct" answers, there is no single "best" method of encoding texts.

Summary (contd.)

- When presuming investigator expertise, respondents' meanings for their words are ignored. Although this approach lends itself to automation, it is restricted to inferences relevant to the investigator's perspective.
- When presuming respondent expertise, texts may afford data either on these respondents' perspectives or on facts they have experienced. (One of these always introduces noise into findings about the other.)

Summary (contd.)

- **Inferences about semantic relations are only legitimate if these relations have been encoded in one's data.**

A recommendation

- **Since the community of scholars who analyze texts is so small, why not give access to the verbatim transcripts only to the subset of these scholars who have interest in ISR national survey data?**

- **Respondents' anonymity could be ensured via investigators' signed "Pledge to Safeguard Respondent Privacy" (along with promises to hand over their first born children if this anonymity is breached).**